Discretization-Based Ensemble Model for Robust Learning in IoT

Anahita Namvar[1](✉), Chandra Thapa[2], and Salil S. Kanhere[1]

[1] UNSW Sydney, Kensington, NSW 2052, Australia
a.namvar@student.unsw.edu.au, salil.kanhere@unsw.edu.au
[2] Data61 Marsfield, Sydney, NSW 2122, Australia
chandra.thapa@data61.csiro.au

Abstract. The rapid proliferation of Internet of Things (IoT) devices has introduced new challenges in network management and security. While machine learning models hold promise for identifying these devices, they remain vulnerable to adversarial attacks, undermining their accuracy and reliability. This paper addresses the need for robust IoT device identification by proposing a novel approach: a discretization-based ensemble stacking technique. This method harnesses both the protective properties of discretization and the generalization benefits of ensemble methods. Through extensive experimentation, we demonstrate the efficacy of our approach against various adversarial attacks, showcasing its potential to enhance the resilience and accuracy of IoT device identification models in dynamic and uncertain environments.

Keywords: Adversarial Robustness · IoT · Discretization · Ensemble

1 Introduction

The Internet of Things (IoT) has revolutionized the way we interact with our surroundings, connecting a wide range of objects, from household appliances to smartphones and automobiles. However, this connectivity has also introduced several challenges for organizations, particularly with regard to managing access and preventing potentially insecure IoT devices from connecting to their network [1]. To address this challenge, accurate and fast methods are needed to effectively manage the visibility of IoT devices on the network and distinguish compromised devices from legitimate ones.

Machine learning models have emerged as a promising approach for IoT device identification, as they can learn to recognize the unique behavior patterns of each device on the network [2]. However, these models are susceptible to various types of noise and adversarial attacks that can potentially affect their accuracy and performance [3–5]. These attacks can introduce perturbations into input data, leading to misclassifications and undermining the reliability of device identification models. In existing literature, while the application of machine learning for IoT device identification is acknowledged,

A. Zaslavsky et al. (Eds.): MobiQuitous 2023, LNICST 594, pp. 353–367, 2024.
https://doi.org/10.1007/978-3-031-63992-0_23

there is a noticeable gap concerning the development of a comprehensive and effective defence strategy against adversarial attacks in this context. Particularly as a means of enhancing robustness, there is a lack of exploration into discretization and ensemble methods in the realm of IoT device classification. Discretization making the input space less sensitive to small perturbations. Furthermore, Ensemble methods often help in reducing overfitting and improving generalization, which are important factors for model robustness. To address this gap, the following research questions have been formulated:

Research Question 1: How does the incorporation of discretization methods as defence mechanisms for IoT device identification classifiers enhance their security against both white-box and black-box adversarial attacks?

Research Question 2: How does the development of a hybrid approach that effectively integrates discretization and ensemble methods synergistically improve the robustness of IoT device identification classifiers against diverse white-box and black-box adversarial attacks using real-world IoT network traffic data?

We summarized the key contributions of this paper in the following:

Contribution 1: Empirical Validation of Discretization Effectiveness

We practically validate the efficacy of discretization methods as defence mechanisms by showcasing their ability to fortify IoT device identification classifiers against a spectrum of adversarial attacks, pinpointing the most effective techniques. Through rigorous experiments, it establishes these methods as a valuable additional layer of protection, aligning with the broader contributions that enhance IoT security resilience.

Contribution 2: Synergistic Enhancement via Discretization-Based Ensemble Defence

In response to Research Question 2, this paper introduces a novel and comprehensive approach that synergistically leverages discretization and ensemble techniques for enhancing the robustness of IoT device identification classifiers. This contribution advances the field by presenting a holistic defence methodology that combines the strengths of discretization and ensemble methods to create a more resilient defence mechanism. The proposed hybrid approach not only addresses the challenges of both white-box and black-box attacks but also offers a practical solution for real-world IoT network traffic scenarios.

These contributions collectively advance the knowledge in the field of IoT security by demonstrating the potential of discretization methods as defence mechanisms, empirically validating their efficacy, and proposing a sophisticated hybrid approach that showcases the synergy between discretization and ensemble techniques in enhancing the robustness of IoT device identification classifiers.

The rest of this paper is organized as follows: Sect. 2 provides backgrounds and related works. Section 3 presents the details of our research methodology. Section 4 provides our experiments and results, and Sect. 5 concludes the paper.

2 Background and Related Works

2.1 IoT Device Classification and Adversarial Attacks

The vulnerability of machine learning models to adversarial attacks is mostly addressed in the image domain, and less amount of research has been done on the security of machine learning base device identification models. Namvar, et al. [5] presented the vulnerability of device identification models, including Random Forest (RF), Logistic Regression (LR), Feed Forward (FF), and Decision Tree (DT) to adversarial attacks. However, they did not provide any solution to increase the models' robustness against attacks. Singh and Sikdar [6] investigated the effectiveness of generating attacks against a deep learning-based appliance classification model using smart meter data. The study found that adversarial attacks can significantly degrade the accuracy of the model. To investigate the detectability of attacks in the smart home domain, the authors proposed a visualization method that compares the distribution of true data points and adversarial data points. The findings of the study highlight the need for effective defence mechanisms to counter adversarial attacks on deep learning-based appliance classification models in the smart home domain. Meidan, et al. [7] proposed random forest supervised machine learning model to detect suspicious IoT devices connected to organizational networks. The results showed that the algorithm correctly classified, white-listed device types in 99% of cases and detected non-white-listed device types in 96% of test cases. The authors presented the resilience of their method to adversarial attacks. Kotak and Elovici [8] propose a new approach for generating real-time adversarial examples using heatmaps generated by CAM and Grad-CAM++ that can be applied in the computer network domain. Authors revealed that in many cases, an adversarial example created using a heatmap could deceive the payload-based deep IoT device identification solution with up to 100% accuracy.

2.2 Discretization-Based Ensemble as a Defence Strategy

Discretization, also known as quantization or binning, is the process of converting continuous input variables to discrete values by dividing the input range into a fixed number of bins or intervals [9]. Discretization has been introduced for secure deep learning in image datasets [10]. Until recently, discretization robustness was mostly studied in computer vision [11–13]. However, only one research presented discretization as a defence mechanism in protecting tabular data from adversarial attack [14].

There are several methods for discretization [15], some of them are used in this study Equal Width, Equal Frequency, Entropy-based Binning, and Minimum Descriptive Length. These methods divide a feature's range into bins using different approaches.

Figure 1 illustrates a simplified view of data discretization and shows how it improves robust machine learning against adversarial attacks. The White circle in Fig. 1 represents the original input sample, denoted as X_1. The black circle represents the adversarial sample, denoted as $\widehat{X}_1$, which has been maliciously tampered with by an attacker. The attacker introduces a slight perturbation, ε, , to the input data. This perturbation leads to misclassification. However, let us suppose that we have defined bin boundaries $(C_1, C_2, \ldots, C_n)$ after applying a discretization algorithm. The bin boundaries can be

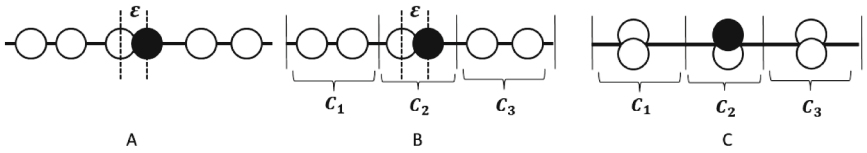

Fig. 1. A simple view of discretization for improving robustness. A. Data before discretization. Continuous variables, and black circle presents an adversarial sample. B. Data is discretized into three equal ranges. All data in the same category have been mapped to an equal value. C. Data after discretization.

seen in Fig. 1. B. After defining bin boundaries, each input data is assigned to the appropriate bin. Finally, the original input data is replaced with the discretized versions in the model. Figure 1.C depicts the discretized samples. Hence, the small changes in the input data cannot fool the model. For example, both X_1 and $\widehat{X}_1$ will be categorized in bin 3, and adversarial attacks cannot simply misclassify input data [16].

Ensemble robustness is another method that can increase the security of machine learning models against adversarial attacks [17–19]. In this study, we illustrate the idea of using an ensemble of different robust discretised classifiers to improve the model's predictive performance in adversarial environment.

3 Methodology

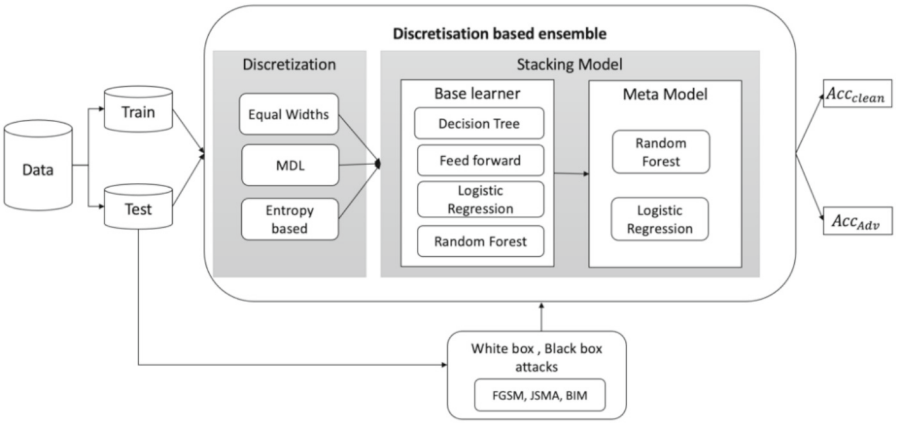

Fig. 2. Discretization-based ensemble robustness.

This section discusses the proposed methodology for the robust learning based IoT device identification mechanism that can improve the resilience of models in an adversarial environment. Figure 2 illustrated the overall methodology of the proposed defence method. We generate FGSM, BIM, and JSMA adversarial attacks in both white and black scenarios. Our defence method specially addresses three different discretization methods of EW, MDL, and EBD collectively enabling the conversion of continuous input

data into discrete representations. Additionally, we utilize stack ensemble to combine different predictions to enhance the model accuracy in adversarial settings.

3.1 Data

In the initial component of the methodology, the study utilizes real-world IoT data to lay the foundation for model development and evaluation. This dataset serves as a crucial source of information, reflecting the complexities of IoT device behaviour in authentic environments. Through rigorous data pre-processing, including cleaning and feature extraction, the dataset is refined for subsequent phases. The process entails a meticulous division into training and testing subsets, enabling the construction and validation of robust machine learning models.

3.2 Discretization

First, we discretized input features into categorical features; we used four state-of-the-art discretization methods including, Equal Frequency (EF), Equal Width (EW), Minimum Description Length (MDL), and entropy-based discretization. These methodologies were chosen due to their ability to effectively discretize continuous data.

In our methodology, each classifier undergoes training with input data that has been transformed using different discretization models. This process results in the creation of multiple base models for each individual classifier, where each base model corresponds to a specific discretization approach. For instance, if we consider three discretization techniques namely, EW, MDL, and EBD applied to the input data, each classifier will be trained separately using these three different discretized datasets. Consequently, for each classifier, a set of three base models emerges, each reflecting the impact of a distinct discretization method.

3.3 Adversarial Attack

In this study, the most state-of-the-art adversarial attacks, including FGSM, BIM, and JSMA, were employed under two scenarios of a white box and a black box to evaluate the robustness of machine learning based IoT device identification models. These attacks can be generated by adding imperceptible perturbations to the input data, which can evade the machine learning models and cause misclassification of IoT devices [5]. For example, an attacker could generate adversarial examples to make a device appear as a legitimate device to avoid detection while it is malicious and performing harmful actions on the network.

3.4 Ensemble

The ensemble method addressed in this study is the stacking ensemble. we address EW, MDL, and entropy discretization techniques since they lead to higher performance on robust accuracy. The stack ensemble model involves two levels of machine learning models: level 0 and level 1. The former consists of single models trained on discretized input features, while the latter is limited to only one meta-model of either RF or LR that combine the outcome of discretized single models.

3.5 Robustness Measures

The "Robustness Measures" component encompasses the quantitative evaluation of model performance under adversarial conditions. Accuracy, a fundamental metric, gauges the model's correctness in predicting device classifications. Meanwhile, we establish the robustness measure as the percentage decrease in model accuracy when comparing the model's performance between a clean and adversarial environment. The Robustness measure (RM) is depicted by Eq. 1.

$$RM = \frac{\|Acc_{Clean} - Acc_{adv}\|_1}{|Acc_{Clean}|} \tag{1}$$

Together, these measures offer a comprehensive insight into the model's capacity to safeguard IoT device identification in dynamic and threat-prone environments.

4 Experimental Results

In this section, we present a case study to demonstrate the practical application and effectiveness of the proposed defence method in IoT device identification. The case study uses a real-world IoT device data set. We focus on the proposed performance metrics to assess the efficacy of our defence mechanism in enhancing the resilience of the ML model against adversarial attacks. The results and analysis highlight the potential of the proposed defence method in securing IoT device classifiers and enhancing their resilience against adversarial threats. The experiments have been run in Python 3.8 on an Intel® Core i5 2.3 GHz CPU, 64 GB of RAM, and running on MAC OS.

4.1 Dataset

We utilized a comprehensive dataset obtained by Sivanathan, et al. [20]. This dataset encompasses a diverse array of network traffic data emanating from 28 distinct IoT devices, encompassing categories such as cameras, lights, motion sensors, appliances, and health monitors. It comprises seven fundamental attributes including traffic flow, flow volume, flow duration, sleep time, NTP interval, DNS interval, and domain count. This dataset forms the cornerstone of our analytical efforts, enabling a holistic exploration of network behaviours exhibited by various IoT devices.

4.2 Adversarial Attack Setting and Evaluation Metric

To evaluate the robustness of the proposed defence mechanism, our study employs three widely recognized adversarial attack techniques: Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and Jacobian-based Saliency Map Attack (JSMA). These attacks are conducted under two distinct scenarios: black box and white box attacks.

In black box attacks, the transferability property of adversarial examples is harnessed to evaluate the resilience of the defence mechanism. We generate adversarial perturbations using substitute models, specifically Logistic Regression (LR) and Feed-forward Neural Network (FF). This approach leads to various combinations, resulting in six different black box attacks: FGSM-LR, FGSM-FF, BIM-LR, BIM-FF, JSMA-LR, and JSMA-FF. For instance, FGSM-LR refers to a black box attack generated using the substitute model of LR. Similarly, FGSM-FF signifies a black box attack created via the substitute model of FF. A complete list of the different combinations of attack and substitute models can be found in Table 1. Such black box attack methods are likely to effectively fool the model due to the transferability property [21, 22]. These adversarial inputs then transferred to the target IoT device classifiers, including Random Forest (RF), Decision Tree (DT), LR, and FF.

In the white box attack setting, the adversary possesses complete knowledge of the learning model, its parameters, and training data. We exclusively employ white box attacks on differentiable classifiers, specifically LR and FF. For non-differentiable classifiers, such as DT and RF, only black box attacks are implemented.

Adversarial examples are crafted using the default settings in IBM's Adversarial Robustness Toolbox [23]. The FGSM attack parameter ϵ is set to o.o1. For a BIM attack, parameters ϵ and α are set to 0.01, and 0.001, respectively. Finally, for the JSMA attack, the parameters, Theta, and gamma are set to 0.01, 0.4, respectively.

In terms of the evaluation metric, we introduce the robustness measure (RM) in Sect. 3.5, represented by Eq. 1. A lower value of RM indicates enhanced robustness. Table 1 presents a comprehensive analysis of the model performance under different black box attack scenarios. The Acc_{adv} column indicates the accuracy of models against black-box attacks, while RM is the robustness measure that quantifies the percentage drop in model accuracy when comparing the model's performance in clean and adversarial environments. The outcome of these black box attacks collectively underscores the fact that no model remains impervious to the influence of adversarial attacks.

Upon analysing the white box attack results presented in Table 2, it becomes evident that the susceptibility of models to adversarial perturbations is not exclusive to black box scenarios. Models such as Logistic Regression and Feed Forward exhibit notable accuracy reductions under various white box attacks, including FGSM, JSMA, and BIM. For instance, Logistic Regression demonstrates a considerable decrease in accuracy across all white box attacks, particularly with FGSM where accuracy drops to 60%. Feed Forward also showcases variations in accuracy reductions depending on the specific attack method.

The results highlight that no model is immune to the impact of adversarial attacks. Under various black and white attacks techniques, including Fast Gradient Sign Method (FGSM) and Jacobian-based Saliency Map Attack (JSMA), accuracy reductions are observed across the board. Some models exhibit more pronounced declines in accuracy and robustness margins than others, indicating varying degrees of susceptibility. While the severity of vulnerability varies depending on the specific attack and model combination, it can be observed that all models experience compromised performance.

Table 1. Black box attacks

Model	Black box Attack	Adv_acc	RM
Decision Tree	FGSM LR	75.78	23.77
	JSMA LR	75.67	23.88
	BIM LR	75.86	23.69
	JSMA FF	74.37	25.19
	FGSM FF	78.11	21.43
	BIM FF	76.61	22.93
LR	JSMA FF	65.21	29.88
	FGSM FF	78.13	15.98
	BIM FF	73.83	20.60
Random Forest	FGSM LR	64.82	35.12
	JSMA LR	73.92	26.01
	BIM LR	66.11	33.83
	JSMA FF	71.52	28.40
	FGSM FF	73.88	26.05
	BIM FF	70.44	29.49
FF	FGSM LR	79.42	13.74
	JSMA LR	65.76	28.58
	BIM LR	78.50	14.74

Table 2. White Box Attacks

Model	Black box attack	Adv_acc	RM
LR	FGSM	60.10%	35.37%
	JSMA	64.82%	30.29%
	BIM	64.98%	30.12%
FF	FGSM	76.00%	17.45%
	JSMA	62.29%	32.34%
	BIM	80.29%	12.79%

4.3 Discretization

In this section, we thoroughly examine the evaluation of how different discretization methods impact the robustness of IoT device classifiers in the adversarial environment. We consider four IoT device identification models namely, Random Forest, Decision

Tree, Logistic Regression, and Feedforward Neural Network. We employ several discretization methods as detailed in Sect. 3.2, including Equal Frequency (EF), Equal Width (EW), Minimum Description Length (MDL), and entropy-based discretization. We examine the models under both white box and black box attack scenarios to comprehensively gauge their resilience. We assess the classifier's robustness to adversarial attacks in terms of accuracy, and RM metric.

First, we show the accuracy of models when the models are not exposed to adversarial manipulation (see Table 3). Original column presents the performance of models in a clean environment where there is no discretization or ensemble mechanism involved. We use this information as a baseline for evaluating the performance of robustness mechanisms. Columns EF, EW, MDL, EBD, EN_LR, and EN_RF correspond to equal frequency, equal width, MDL, entropy-based discretization, discretization-based ensemble with LR meta model, and discretization-based ensemble with RF meta model, respectively, showcasing the performance of models in a clean environment.

Table 3. Models' performance in clean environment

Model	original	EF	EW	MDL	EBD	EN-LR	EN-RF
DT	99.41	50.67	93.61	84.15	85.42	96.27	96.30
RF	99.90	50.91	99.71	99.88	99.80	99.70	99.71
LR	92.99	32.28	99.68	99.88	99.82	99.67	99.59
FF	92.07	50.26	99.21	99.80	99.71	99.60	99.55

Table 4. Black box attacks

Model	Black box attack	EW		MDL		EBD		No Discretisation	
		Adv_acc	RM	Adv_acc	RM	Adv_acc	RM	Adv_acc	RM
Decision Tree	FGSM LR	86.84	7.23	74.51	11.46	73.49	13.97	75.78	23.77
	JSMA LR	85.36	8.81	78.35	6.89	78.41	8.21	75.67	23.88
	BIM LR	87.10	6.95	74.53	11.43	73.69	13.73	75.86	23.69
	JSMA FF	81.34	13.11	71.18	15.41	74.06	13.30	74.37	25.19
	FGSM FF	87.27	6.77	74.35	11.65	75.20	11.96	78.11	21.43
	BIM FF	86.28	7.83	71.55	14.97	73.75	13.66	76.61	22.94

(*continued*)

Table 4. (*continued*)

Model	Black box attack	EW		MDL		EBD		No Discretisation	
		Adv_acc	RM	Adv_acc	RM	Adv_acc	RM	Adv_acc	RM
Random Forest	FGSM LR	96.87	2.85	93.83	6.06	94.91	4.90	64.82	35.12
	JSMA LR	98.57	1.14	98.49	1.39	98.69	1.11	73.92	26.01
	BIM LR	97.04	2.68	96.34	3.54	96.36	3.45	66.11	33.82
	JSMA FF	97.18	2.54	94.77	5.12	96.22	3.59	71.52	28.41
	FGSM FF	96.24	3.48	97.91	1.97	96.75	3.06	73.88	26.05
	BIM FF	96.42	3.30	94.58	5.31	93.25	6.56	70.44	29.49
LR	JSMA FF	95.91	3.78	92.91	6.98	92.89	6.94	65.21	29.87
	FGSM FF	94.77	4.93	96.69	3.19	95.20	4.63	78.13	15.98
	BIM FF	95.24	4.45	93.72	6.17	92.29	7.54	73.83	20.60
FF	FGSM LR	92.54	3.81	91.97	7.85	87.57	12.18	79.42	13.74
	JSMA LR	90.97	5.45	96.36	3.45	92.07	7.66	65.76	28.58
	BIM LR	92.52	3.84	91.95	7.87	87.20	12.55	78.50	14.74

It is apparent that the EF discretization method has led to a significant decrease in the accuracy of the models. The original models have high accuracy, while the model built using the EF discretization method has much lower accuracy. This indicates that the discretization process has negatively impacted the ability of the models to accurately predict outcomes. This is mainly due to the distribution of the data, as the model cannot discretize the data into bins with equal frequency when the frequency in some bins is very high. The maximum number of bins for equal frequency is 17, and the model is unable to increase this limit. Given the poor performance of EF approach in a clean environment, we will not evaluate its robustness capability in the next steps. However, the results indicate that the remaining methods, namely EW, MDL, EBD, EN-LR, and EN-RF accurately identify IoT devices in a clean environment. Consequently, we will evaluate the performance of these models in an adversarial environment. Table 4 presents the accuracy of different discretized models when subjected to different black box attacks.

while RM is the robustness measure. Most of the discretized models exhibit a high level of adversarial robustness, as evidenced by their accuracy in an adversarial environment, compared to their lower accuracy when no discretization mechanism is used. EW is proving to be a robust technique for all classifiers, as it achieves high accuracy in adversarial environments and exhibits less degradation in accuracy when exposed to adversarial attacks (low RM measure). For example, in the Decision Tree model, the Adv_acc for FGSM LR attack is 86.84% in the EW model, whereas it is only 75.78% without discretization. Similarly, in the Random Forest model, the Adv_acc for JSMA LR attack is 98.57% in the EW model, whereas it is only 73.92% in the no discretization model. This can be attributed to its ability to handle data with varying distributions and to its effectiveness in reducing the impact of outliers. EW discretizes the variable into a smaller number of categories; hence, the noise and variability can be reduced, and the model can learn more general and robust patterns. Additionally, equal-width discretization is a simple and easy-to-implement technique that requires minimal computational resources. Overall, the strong performance of equal-width discretization suggests that it is a promising approach for achieving adversarial robustness in this domain.

Table 5. White box attacks

Model	White box attack	EW		MDL		EBD		No Discretisation	
		Adv_acc	RM	Adv_acc	RM	Adv_acc	RM	Adv_acc	RM
LR	FGSM	94.79	4.91	96.50	3.38	94.44	5.39	60.10	35.37
	JSMA	96.50	3.19	98.22	1.66	97.20	2.62	64.82	30.29
	BIM	94.83	4.87	96.46	3.42	94.32	5.51	64.98	30.12
FF	FGSM	93.30	5.96	93.01	6.80	90.39	9.35	76.00	17.45
	JSMA	88.12	11.18	93.17	6.64	85.24	14.51	62.29	32.34
	BIM	94.89	4.35	91.23	8.59	87.84	11.90	80.29	12.79

Results demonstrate the effectiveness of the MDL methodology in increasing the resilience of random forest, feedforward, and logistic regression models to the black box. The MDL methodology achieves this by minimizing the description length of the model, which encourages the selection of simpler and more interpretable models that are less prone to overfitting. In addition to improving model robustness, the MDL methodology provides a principled approach for selecting the best model among a set of candidates based on the tradeoff between model accuracy and complexity. Overall, the MDL methodology is a promising approach for achieving robustness in IoT device classification. The EBD discretization method has emerged as an effective way to increase the robustness of machine learning models to black box attacks, except for the decision tree model. For example, in the case of the random forest model, the EBD approach improved the adversarial accuracy against the JSMA LR attack from 73.92% to 98.69% and against the FGSM FF attack from 73.88% to 96.75%. Overall, the effectiveness of the EBD approach in improving the robustness of models against black-box attacks

appears to be model-dependent. Further analysis to understand the factors that affect the effectiveness of EBD will be left for future research.

Table 5 illustrates the adversarial robustness of discretized models to white-box attacks. All classifiers without input discretization are vulnerable to such attacks. In contrast, IoT device classifiers trained on discretized input data generated by EW, MDL, EBD show less vulnerability to adversarial attacks as their Adv_acc is much greater than that of models without discretization. Also, discretized models present lower RM compared to models with no discretization, which confirms the effectiveness of input discretization with EW, MDL, EBD for creating a robust IoT device identification model in a white box scenario.

4.4 Discretization-Based Ensemble Defence Mechanism

Table 6. Discretization based ensemble-black box attacks

Model	Black box attack	EN-LR		EN-RF		Original model (no robustness)	
		Adv_acc	RM	Adv_acc	RM	Adv_acc	RM
Decision Tree	FGSM LR	91.70	4.75	91.86	4.61	75.78	23.77
	JSMA LR	95.13	1.18	95.83	0.49	75.67	23.88
	BIM LR	93.07	3.32	91.02	5.48	75.86	23.69
	JSMA FF	91.15	5.32	92.54	3.90	74.37	25.19
	FGSM FF	94.50	1.84	92.70	3.74	78.11	21.43
	BIM FF	92.80	3.60	90.17	6.37	76.61	22.93
Random Forest	FGSM LR	97.77	1.94	97.34	2.38	64.82	35.12
	JSMA LR	99.36	0.34	99.59	0.12	73.92	26.01
	BIM LR	97.58	2.13	97.54	2.18	66.11	33.83
	JSMA FF	97.52	2.19	97.28	2.44	71.52	28.40
	FGSM FF	98.20	1.50	98.16	1.55	73.88	26.05
	BIM FF	96.58	3.13	96.55	3.17	70.44	29.49
LR	JSMA FF	95.95	3.73	96.71	2.89	65.21	29.88
	FGSM FF	97.89	1.79	96.49	3.11	78.13	15.98
	BIM FF	96.40	3.28	96.54	3.06	73.83	20.60
FF	FGSM LR	96.03	3.58	95.68	3.89	79.42	13.74
	JSMA LR	96.40	3.21	96.18	3.39	65.76	28.58
	BIM LR	96.39	3.22	95.98	3.59	78.50	14.74

We evaluated the effectiveness of the proposed discretization-based ensemble robustness methodology in enhancing the robustness of IoT device classifiers against adversarial

attack. Two discretisation-based ensembles of EN-LR, EN-RF were built in this study. The ensemble method employed was a stacking method using the Scikit-learn Python library. We achieved this through the aggregation of individually trained models, each employing distinct discretization strategies on input features, resulting in diverse base classifiers for the same IoT device classification; these base models were then used to build a stack ensemble, ultimately yielding a robust classifier tailored to the base IoT device classifier. The main difference between EN-LR and EN-RF is their choice of meta-model. EN-LR uses LR as its meta model, while EN-RF uses RF as its meta-model. For example, to build EN-LR for DT model, The DT with EW discretisation, DT with MDL, and DT with EBD were used as base learners and LR as a meta-model.

Table 6, and Table 7 illustrate the performance of our models in black-box and white-box scenarios, respectively. Each table shows the adversarial accuracy and robustness measures of the models against different adversarial attacks EN-LR, EN-RF, and original models without robustness methodology. Looking at Table 6, we can see that across various black box attacks, both EN-LR and EN-RF consistently outperform the original models in terms of adversarial accuracy. For instance, within the context of DT model, the application of the EN-LR discretization-based ensemble robustness methodology yields a remarkable improvement in model accuracy, achieving accuracy of 91.70% against the FGSM-LR adversarial attack. This achievement is particularly noteworthy when compared to the original DT model's accuracy of 75.78% when exposed to the same adversarial attack. Furthermore, the robustness measures (RM) for EN-LR and EN-RF are generally lower than those of the original models. For instance, analysing the RF classifier robustness to JSMA-LR black box attack, the EN-LR robustness method achieves an RM of 0.34%, EN-RF an RM of 0.12%, whereas the original model's RM is significantly higher at 26.01%. Both robustness approaches exhibit notable improvements in adversarial accuracy and reduced robustness measures. Table 7 reveals insights into the effectiveness of EN-LR, EN-RF discretization based ensemble approaches in enhancing the robustness of the models against white box attacks. The EN-LR and EN-RF models consistently demonstrate higher adversarial accuracy compared to the original models across various white box attacks. For instance, LR model using EN-LR robustness methodology achieves an accuracy of 96.08% against FGSM white box

Table 7. Discretization based ensemble - white box attacks

Model	Black box attack	EN-LR		EN-RF		Original model (no robustness)	
		Adv_acc	RM	Adv_Acc	RM	Acc_Adv	RM
LR	FGSM	96.51%	3.17%	96.18%	3.42%	60.10%	35.37%
	JSMA	98.71%	0.96%	98.94%	0.65%	64.82%	30.29%
	BIM	96.39%	3.29%	95.13%	4.48%	64.98%	30.12%
FF	JSMA	94.62%	5.00%	94.28%	5.29%	62.29%	32.34%
	FGSM	93.50%	6.12%	93.42%	6.16%	76.00%	17.45%
	BIM	94.39%	5.23%	95.23%	4.34%	80.29%	12.79%

attack, outperforming the original model's accuracy of 60% for same attack. It clearly highlights the improvement achieved by the EN-LR approach.

5 Conclusion

In this paper, we introduced a discretization-based stack ensemble methodology to enhance the resilience of machine learning-driven IoT device identification models against adversarial attacks. Our findings highlight the efficacy of discretization as a strategy to improve model robustness across white and black box attack scenarios. Our research's generality and adaptability make it applicable to diverse IoT datasets, increasing its relevance in cybersecurity. In the future, we will delve into the practical implementation of our approach within a new domain. By exploring the application of our method in other contexts, we aim to solidify the real-world utility further.

References

1. Al-Garadi, M.A., Mohamed, A., Al-Ali, A.K., Du, X., Ali, I., Guizani, M.: A survey of machine and deep learning methods for internet of things (IoT) security. IEEE Commun. Surv. Tutor. **22**(3), 1646–1685 (2020)
2. Cvitić, I., Peraković, D., Periša, M., Gupta, B.: Ensemble machine learning approach for classification of IoT devices in smart home. Int. J. Mach. Learn. Cybern. **12**, 1–24 (2021)
3. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access **6**, 14410–14430 (2018). https://doi.org/10.1109/ACCESS.2018.2807385
4. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: Attacks and defenses for deep learning. IEEE Trans. Neural Netw. Learn. Syst. **30**(9), 2805–2824 (2019). https://doi.org/10.1109/TNNLS.2018.2886017
5. Namvar, A., Thapa, C., Kanhere, S.S., Camtepe, S.: Evaluating the security of machine learning based IoT device identification systems against adversarial examples. In: Hacid, H., Kao, O., Mecella, M., Moha, N., Paik, H. (eds.) Service-Oriented Computing: 19th International Conference, ICSOC 2021, Virtual Event, November 22–25, 2021, Proceedings, pp. 800–810. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-91431-8_57
6. Singh, A., Sikdar, B.: Adversarial attack for deep learning based IoT appliance classification techniques. In: 2021 IEEE 7th World Forum on Internet of Things (WF-IoT) , pp. 657–662. IEEE (2021)
7. Meidan, Y., et al.: Detection of unauthorized IoT devices using machine learning techniques. arXiv preprint (2017)
8. Kotak, J., Elovici, Y.: Adversarial attacks against IoT identification systems. IEEE Internet Things J. **10**(9), 7868–7883 (2023). https://doi.org/10.1109/JIOT.2022.3229906
9. Brownlee, J.: Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. In: Machine Learning Mastery (2020)
10. Buckman, J., Roy, A., Raffel, C., Goodfellow, I.: thermometer encoding: one hot way to resist adversarial examples. In: International Conference on Learning Representations (2018)

11. Sharmin, S., Rathi, N., Panda, P., Roy, K.: Inherent adversarial robustness of deep spiking neural networks: effects of discrete input encoding and non-linear activations. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX, pp. 399–414. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58526-6_24
12. Panda, P., Chakraborty, I., Roy, K.: Discretization based solutions for secure machine learning against adversarial attacks. IEEE Access **7**, 70157–70168 (2019)
13. Lal, S., et al.: Adversarial attack and defence through adversarial training and feature fusion for diabetic retinopathy recognition. Sensors **21**(11), 3922 (2021)
14. Kireev, K., Kulynych, B., Troncoso, C.: Adversarial robustness for tabular data through cost and utility awareness. arXiv preprint arXiv:2208.13058 (2022)
15. Maslove, D.M., Podchiyska, T., Lowe, H.J.: Discretization of continuous features in clinical datasets. J. Am. Med. Inform. Assoc. **20**(3), 544–553 (2013)
16. Zhou, J., Zaidi, N., Zhang, Y., Li, G.: Discretization inspired defence algorithm against adversarial attacks on tabular data. In: Gama, J., Li, T., Yang, Y., Enhong Chen, Y., Zheng, F.T. (eds.) Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16–19, 2022, Proceedings, Part II, pp. 367–379. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-05936-0_29
17. Kurakin, A., et al.: Adversarial attacks and defences competition. In: Escalera, S., Weimer, M. (eds.) The NIPS '17 Competition: Building Intelligent Systems, pp. 195–231. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-94042-7_11
18. Mohanty, H., Roudsari, A.H., Lashkari, A.H.: Robust stacking ensemble model for darknet traffic classification under adversarial settings. Comput. Secur. **120**, 102830 (2022)
19. Li, D., Li, Q.: Adversarial deep ensemble: evasion attacks and defenses for malware detection. arXiv preprint arXiv:2006.16545 (2020)
20. Sivanathan, A., et al.: Classifying IoT devices in smart environments using network traffic characteristics. IEEE Trans. Mob. Comput. **18**, 1745–1759 (2018)
21. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P) , pp. 372–387. IEEE (2016)
22. Papernot, N., McDaniel, P., Goodfellow, I.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint (2016)
23. Nicolae, M.-I., et al.: Adversarial Robustness Toolbox v1. 0.0. arXiv preprint arXiv:1807.01069 (2018)