2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 11-14, 2020. Toronto, Canada

A Technique for Generating a Botnet Dataset for Anomalous Activity Detection in IoT Networks

Imtiaz Ullah, Qusay H Mahmoud
Department of Electrical, Computer and Software Engineering
Ontario Tech University,
Oshawa, ON, L1G 0C5 Canada
{imtiaz.ullah, qusay.mahmoud}@ontariotechu.net

***Abstract*—In recent times, the number of Internet of Things (IoT) devices and the applications developed for these devices has increased; as a result, these IoT devices are targeted by many malicious activities that cause potential damage in many smart infrastructures. A technique is required to appropriately classify anomalous activities to minimize the impact of these activities. The IoT networks are difficult to analyze and test because of the lack of sufficient well-structured IoT datasets for anomaly-based intrusion detection. In this paper, we present a technique we have used to generate a new Botnet dataset, from an existing one, for anomalous activity detection in IoT networks. The new IoT botnet dataset has a wider network and flow-based features. A flow-based Intrusion Detection System (IDS) can be analyzed and tested using flow-based features. Finally, we use different machine learning methods to test the accuracy of our proposed dataset. We also test the accuracy of our proposed dataset through various feature correlation and the methodology for recursive feature elimination. Our proposed IoT botnet dataset provides a ground to analyze and evaluate anomalous activity detection model for IoT networks. We have shared the newly generated Botnet dataset publicly, and a link is provided in this paper.**

Keywords—Internet of Things, Intrusion detection, IDS Dataset, DDoS, DoS, flow-based intrusion detection, anomaly detection system, Infiltration, IoT Dataset, cybersecurity.

I. INTRODUCTION

The IoT, so-called smart devices, transformed the device's interaction methodology at home, work, or work area. The IoT devices are generally resource-constrained, battery-powered, and connected to the Internet. The Internet connection makes IoT networks more exposed to several threats. The anomaly-based IDS can detect attacks at the network level. The intrusion detection implemented via machine learning techniques makes it possible to detect a diversity of existing attacks as well as a variety of new attacks. Reliability is an essential requirement for an anomaly detection system. An IDS is considered more accurate if it achieves a higher detection rate and a low false-positive rate. The primary objective of anomaly-based IDS is to achieve a high detection rate. Security becomes a significant and critical part of the IoT networks due to the increasing number of cyberattacks on IoT networks. An IDS is useful only if it detects novel attacks. Anomaly-based detection procedures can be used in medical diagnoses, fraud detection, and new topic detection text mining.

An anomaly in the smart home can put human life in danger if the anomaly is not detected on time. If a patient sensor in a smart home sends false information to a caregiver or if the sensor does not inform an emergency worker on time or an attacker overdoses, a patient via an actuator can put a person's life in danger. A sensor monitoring medical control environment may cause severe injuries to human life, or an attacker can disable a smart alarm system to rob a house or a building. A motion and camera sensor malfunctioning allow an intruder to get access to a smart home or smart building, e.g., a DoS attack knocked out a CCTV camera in smart infrastructure, and the camera will not report data. An IoT device can be hijacked to launch a DDoS attack, or a smart home device can be hacked to cause the disorder. An attacker can also spoof an IoT device to extract cryptographic information and subsequently access a smart system using the identity of the hacked device to spy private audio or video conversation. Hackers can deactivate the safety features of connected cars to put human life in danger.

IoT networks use different connectivity domains in smart infrastructure, e.g., IoT covers fog computing, cloud computing, sensor networks, traditional Internet, and mobile networks. Therefore, conventional security techniques may not be appropriate for IoT networks. The attackers used more enhanced methodologies to introduce more severe malicious activities very quickly in IoT networks. The attacker needs very little technical knowledge of IoT networks to launch these attacks [1]. Presently, a successful attack has a widespread impact on IoT resources, while the time required for these attacks becomes tremendously shorten. It is predicted that the attackers will be able to affect the IoT networks in a matter of seconds very soon [1]. Fig. 1 shows the impact of time and possible destruction of cyber threats. Companies, hospitals, government, and public services were affected by ransomware attacks in many countries in recent years. The attackers were able to conceal public and private data at a very high propagation speed, industry paying millions of dollars to unlock encrypted files. Anomaly-based approaches for IoT networks still an immature technology in state-of-the-art IDS solutions and commercial tools.

IDS use many types of input data for its functionality. The input data include system logs, network traffics, application logs, binary or raw alerts, event traces, and threat information. Disk, memory, packet, function, code, or conventional logging are the monitoring methods used by an IDS to collect information dynamically or statically from an application or system. An IDS can analyze and detect malicious activity at the network level, host level, or both network and host level. An intrusion can be detected either statically or dynamically. Sometimes the data gathering, and the detection processes are realized separately,

978-1-7281-8526-2/20/$31.00 ©2020 IEEE

e.g., data may be gathered dynamically but subsequently analyzed statically. Conventional detection methods are pattern-based, anomaly-based, ontology-based, graph-based, and similarity-based [2]. A simple correlation system uses several methods to improve the detection accuracy. The analysis of malicious activity is not only limited to malware but can be extended to the behavior or contextual analysis. Intrusion detection technique can be real-time, delayed fixed interval, or request initiated by user command or upon detection of an event. The data processing technique can be local, centralized, or distributed to process collected or store data [2].

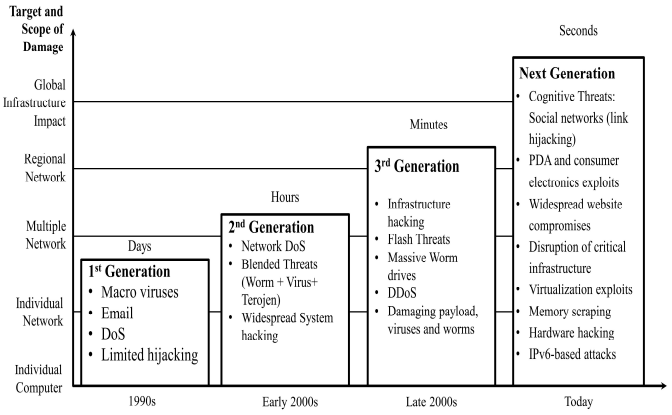

Fig.1. The Impact of Time and Possible Destruction of Cyber Threats [1]

Supervised and unsupervised methods may be used to produce information from the data obtained. The supervised learning technique uses an existing anomalous data model to define the inconsistency to a reference point. In comparison, the anomalous behavior is decided through an unsupervised learning strategy by inferential learning to make a decision. Popular machine learning techniques are Support Vector Machine, Markov-based, Grammar-based, Neural network, Bayesian, Decision tree, Nearest prototype, Hierarchical, and Outlier-based identified techniques for classification and clustering [2].

The remainder of this paper is organized as follows: The related work is discussed in section II and preceded in section III by our proposed IoT botnet dataset system. The analysis of the results is presented in Section IV. Finally, we conclude the paper in Section V and propose ideas for future work.

II. Related Work

IoT has recently gained considerable attention within the information technology industry. Smart homes, smart cities, healthcare, smart manufacturing are general IoT application areas. IoT vulnerability increased expressively as integrated IoT networks grew in demand and development. A well-structured IoT dataset for anomaly-based intrusion detection is not available. Therefore, it is challenging to analyze and evaluate various techniques for anomaly detection in IoT networks. KDD99 is the most popular dataset used for the evaluation of anomaly-based intrusion detection systems. The KDD99 is the most extensively used dataset for intrusion detection. However, the KDD99 dataset is widely criticized because the dataset is biased towards the attack instances. The KDD99 dataset consists of 80% of attack traffic and 20% of normal network traffic. Ghorbani et al. [3] investigated the KDD99 dataset to remove redundant records from training and testing sets. The new dataset is named the NSL-KDD dataset.

Ring et al. [4] conducted a comprehensive study of the intrusion detection datasets. They compared various intrusion detection datasets and analyze these datasets based on the following five criteria: general information, data volume, nature of the data, recording environment, and evaluation. Mirsky et al. [5] developed Kitsune, an online lightweight autoencoder IDS. The Kitsune used an unsupervised learning approach to extract network features dynamically from network traffic. The Kitsune IDS was evaluated via video surveillance IoT network. The botnet was widely used in numerous cyber-attacks and triggered serious threats to several organizations. Botnet detection is a difficult issue for security professionals in the computer network [6]. Koroniotis et al. [7] developed an IoT botnet dataset. They considered several botnet scenarios to capture anomaly network traffic and normal network traffic. They considered several botnet scenarios to collect anomaly network traffic and normal network traffic, but their dataset has a limited number of general and flow-based features. New techniques and detection algorithms for IoT networks needed a well-structured dataset. In earlier work [8], we generated a new dataset named IoTID20 for anomalous activity detection in IoT networks. The IoTID20 dataset provides a large number of general network features and flow-based features. The IoTID20 dataset replicates modern IoT network communication traffic and is freely available. The IoTID20 dataset provides a framework for the implementation of new intrusion detection strategies in IoT networks.

Several researchers proposed different solutions to study the behavior of HTTP based botnet attacks. Saad et al. [9] suggested a technique for Peer-to-Peer botnet discovery to recognize botnet before the attack was initiated. For their experiments, they used five machine-learning strategies and 17 network features. Their framework used traffic behavior analysis to detect botnet during a command and control stage. A primary requirement for the development of a dataset is to generate real network traffic. The Canadian Institute of Cybersecurity(CIC) profile human behavior interaction to generate a new dataset for intrusion detection [10]. The new dataset is more reliable and contains the latest network attacks and benign network traffic. Thamilarasu and Chawla proposed an IDS for IoT networks [11]. They evaluate their framework via real network traffic. The goal of their approach is to detect IoT malicious activity in real-time. They used deep neural network (DNN) with three hidden layers to construct their proposed framework.

Distributed denial of service attacks is a severe threat to computer networks. The Canadian Institute of Cybersecurity has developed a new DDoS dataset [12] to remove weaknesses in the existing datasets [10],[13]. The new dataset provided a wide range of flow-based and general network features. Benign profile agents are used to generate realistic network traffic. Benign profile agents abstract the behavior of humans via SSH, Email, HTTP, HTTPS, and FTP Protocols. The DDoS dataset

is generated via conventional network communication; therefore, the DDoS dataset cannot be used to analyze and evaluate IoT networks. Meidan et al. [14] used deep autoencoder to detect IoT botnet attacks. The deep autoencoder framework was evaluated via Mirai and Bashlite IoT botnets. They used nine IoT devices to develop the IoT botnet dataset for their proposed framework. A device which has more functional capabilities received more FPR. Johansson and Olsson modify snort to develop an IDS for IoT networks [15]. Their framework decreased energy consumption and accelerated the network traffic analysis of snort. In earlier work [16]-[18], we proposed a two-level IDS model for IoT networks. The model was evaluated via CICIDS 2017 and UNSW-NB15 datasets. We used RFE for feature selection and SMOTE, and ENN was used to balance the dataset. Our proposed model achieved a very satisfactory detection rate. Chauhan et al. [19] studied various classification techniques for intrusion detection. The purpose of this study is to find the best classification algorithm for intrusion detection. Unsupervised machine learning algorithms are not common for intrusion detection. Duque and Omar [20] proposed a model for intrusion detection via an unsupervised machine learning technique using K-means clustering. Their proposed model achieved better accuracy and low false positive. Their model successfully evaluates and identifies the behaviors and signature of attacks.

III. THE PROPOSED IOT BOTNET DATASET

IDS becomes an essential part of modern computer networks to detect malicious activities in real-time. An IDS observe, identify, and outline malicious events or policy destruction in the computer system or a computer network. With the dramatic increase in computer attacks, the IDS becomes an essential element of a computer network [2]. Privacy is a key challenge associated with intrusion detection datasets, which limit the availability of intrusion detection datasets to the public. Table I shows a list of publicly available datasets for anomaly detection. MIT Lincoln LAB sponsored the first IDS evaluation event in 1998 to develop an intrusion detection dataset. The DARPA data collected in a simulated environment. There are two parts of the DARPA 1998 dataset, real-time evaluation and an offline evaluation. The primary purpose of this event to generate a dataset to evaluate the detection capability of an IDS. The dataset was collected in the form of TCP/IP dump files for seven weeks. The first two weeks of data are attack free normal data, which is suitable for training a machine learning algorithm. The anomalous portion of the dataset was generated for five weeks in a simulated environment.

The KDD99 dataset was generated from data captured in the DARPA 1998 event. Lee and Stolfo [21] extract the KDD99 dataset features from the raw DARPA TCP/IP dump files. The KDD99 dataset consists of two weeks of attack-free data and five weeks of attack data. The main categories of the KDD99 dataset are U2R, R2L, Probe, DOS, and Normal. There are different types of attacks under each category. The training dataset contains 24 attack categories, while the testing dataset consists of 14 attack categories. The testing data contains new attacks that are not available in training data, which allow measuring the IDS capability towards unknown or new attacks. The KDD99 dataset is a comprehensively biased dataset toward the attack instances because the KDD99 dataset consists of 80% of attack traffic and 20% of normal network traffic. In comparison, the real network contains 99.99% of normal network traffic. The KDD99 dataset's frequent classes are DoS and Probe, while less frequent classes are U2R and R2L. The redundant records in training and testing set produce bias results for Normal and DoS classes. NSL-KDD dataset removed these deficiencies of the KDD99 dataset. Ghorbani et al. [3] eliminate 78 % redundant instances from the training set and 75 % redundant instances from the testing set of the KDD99 dataset. These redundant instances generate biased results in training and testing a machine learning algorithm. The New dataset was named NSL-KDD [3]. Due to the lack of the publicly available intrusion detection datasets, the NSL-KDD dataset was considered an important dataset for valuing intrusion detection framework.

TABLE I. INTRUSION DETECTION DATASETS

	Dataset Name	Description
1	DARPA 98-99	Lincoln Laboratory 1998-99
2	KDD99	KDD99 Dataset
3	DEFCON	DEFCON-10 2002
4	CAIDA	Center of Applied Internet Data Analysis
5	LBNL	Lawrence Berkeley National Laboratory
6	CDX	United State Military Academy
7	Kyoto	Kyoto University
8	Twente	University of Twente
10	UMASS	University of Massachusetts
11	ISCX2012	University of New Brunswick
12	ADFA	University of New South Wales
13	UNSW-NB15	University of New South Wales
14	CICIDS2017	University of New Brunswick
15	BoT-IoT	BoT IoT Dataset
16	IoT Botnet	IoT Botnet Attack Dataset
17	IoTID20	IoT Intrusion Detection Dataset 2020

The dataset proposed in this paper was adapted from [7] Pcap files. The new dataset is named IoT Botnet dataset, which can be accessed at [22]. Our proposed IoT Botnet dataset contains more general network features and more flow-based network features. The components of the IoT botnet dataset testbed are simulated IoT services, network platforms, and ISCXFlow meter. The network platform comprises of normal and attacking virtual machines. The IoT services are simulated through the node-red tool. The network features were extracted using an ISCXFlow meter application [23]. The testbed of our proposed dataset consists of VMs that are connected to LAN and WAN [7]. VMs linked to the Internet through the PFSense machine. A packet-filtering firewall with two interface cards was used to ensure the validity of the labeling process of the dataset. The VMs network consists of Ubuntu tap machine, Metasploitable, Ubuntu Mobile, Windows 7, and four Kali Linux machines. IoT services are implemented in the Ubuntu server to mimic a real IoT network, and the Kali Linux machines were used as attacking systems. The normal network traffic is generated via the ostinato tool [7]. The following IoT services were implemented in the dataset testbed environment: Motion-activated lights, Smart fridge, Smart thermostat,

Remotely activated garage door and Weather station. The simulated IoT services are connected to the MQTT broker in the Ubuntu server as well as to the AWS IoT hub. A typical smart home environment was created with five IoT devices operated locally and connected to the cloud infrastructure via the node-red tool for generating normal network traffic. MQTT protocol is used to transfer messages from IoT devices to the cloud. Fig. 2 shows the attack taxonomy of the proposed dataset. We used the ISCXFlow meter to convert Pcap files to CSV files. The ISCXFlow meter extracts 80 statistical network features from Pcap files [23]. The proposed IoT botnet dataset contains 83 network features and three label features. The label features are binary, category, and subcategory. Details of binary, category, and subcategory are presented in Table II. To the best of our knowledge, there are only two IoT botnet datasets available in the literature for evaluating an IDS for IoT networks.

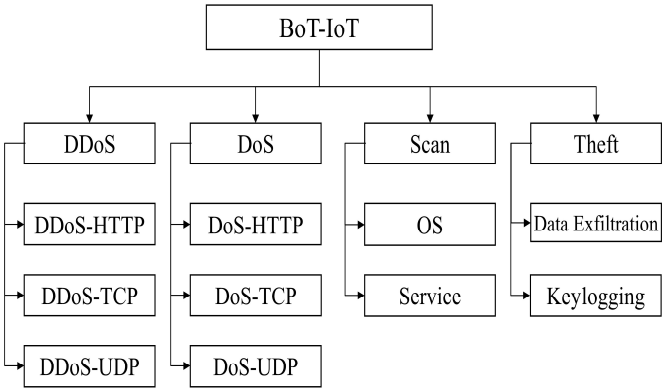

Fig.2. IoT-Botnet Attack Taxonomy

TABLE II. BINARY, CATEGORY, AND SUB-CATEGORY OF IOT BOTNET DATASET

Binary	Category	Subcategory
Normal, Anomaly	Normal, DDoS, DoS, Scan, Theft	DDoS (HTTP, TCP, UDP), DoS (HTTP, TCP, UDP), OS, Services, Data Exfiltration, Keylogging

Koroniotis et al. [7] considered several botnet scenarios to capture anomaly network traffic and normal network traffic, but the dataset has an inadequate number of regular and flow-based features. Meidan et al. [14] used nine IoT devices to develop an IoT botnet dataset. Their dataset has many network features, but most of the features were derived from a limited number of packet features. Another drawback is the absence of flow-based features in the dataset. Our proposed IoT botnet dataset provides many uncorrelated flow-based features and network features as compared to the dataset IoTID20 [8]. Our generated dataset provides a foundation for developing a flow-based intrusion detection model for IoT networks. The IoT botnet dataset required a preprocessing process because the data types and the format of some features are not suitable for machine learning algorithms. We used supervised machine learning algorithms and column normalization techniques to normalize and evaluate the IoT botnet dataset. Accuracy, precision, recall, and F score matrices are used to analyze the IoT botnet dataset.

IV. ANALYSIS

The primary objectives of an IDS are to maximize the true positive and true negative rates and minimize the false positive and false negative rates. For our experiments, we examined and used several classifiers to evaluate our proposed IoT botnet dataset. In this paper, we choose the five most excellent classifiers based on accuracy, precision, recall, and F score for binary, category, and subcategory classes. Five classifiers used in this paper are Logic Regression, GaussianNB, LDA, Ensemble, Decision Tree, Random Forest. Fig. 3 shows the time required for training and testing for different classifiers used in this paper.

Fig.3. Time for Traning and Testing

A. Binary Classification

The binary label classifies the IoT botnet dataset as normal network traffic or malicious network traffic. The performance of all classifiers was outstanding for both normal and anomaly class. We used a K-fold cross-validation test to check the performance and overfitting of different models. The efficiency of these models was tested using cross-validation of 3, 5- and 10-folds. The results of the cross-validation test remain constant for all models. Table III shows the F-score of the 10-fold cross-validation test for binary classification.

B. Category Classification

The Category label organizes the dataset as normal network traffic or any of the following attack categories: DDoS, DoS, Scan, or Theft. Some classifiers perform very well for category label classification, while other classifier's performance was lower for category label classification. A model used a random forest classifier performed very well for all classes of the category of the IoT botnet dataset. Random forest classifier also required less time for training and testing, as shown in Fig. 3. Gaussian NB, LDA, and Logic regression performed poorly for DDoS, DoS, and Scan attacks. The validity of these models has been tested using a 3, 5- and 10-fold cross-validation. Table IV shows the F-score of the 10-fold cross-validation test for category classification.

C. Subcategory Classification

The subcategory label categorizes the IoT botnet dataset as normal network traffic or any of the following attack categories; DDoS-HTTP, DDoS-TCP, DDoS-UDP, DoS-HTTP, DoS-TCP, DoS-UDP, OS, Services, Data Exfiltration, Keylogging. Some classifiers performed very well for subcategory

classification, while other classifier's performance was very poor for some subcategory classification. A model used decision tree classifier performed very well for all classes of subcategories. The decision tree classifier required less time for training and testing, as shown in Fig. 3. The performance of Gaussian NB, LDA, and Logic regression were not satisfactory for DDoS, DoS, and Scan attacks. The effectiveness of these models has been tested using a 3, 5- and 10-fold cross-validation. Table V shows the F score of the 10-fold cross-validation test for subcategory classification.

TABLE III. F SCORE FOR BINARY CLASSIFICATION

	GaussianNB	LDA	Logistic Regression	Decision Tree	Random Forest	Ensemble
Normal	99.99	99.99	99.99	100	100	100
Anomaly	99.99	99.99	99.99	100	100	100

TABLE IV. F SCORE FOR CATEGORY CLASSIFICATION

	GaussianNB	LDA	Logistic Regression	Decision Tree	Random Forest	Ensemble
Normal	95	97	94	99	99.99	99.99
DDoS	45	80	81	99	99.99	99.99
DoS	46	72	59	99	99.99	99.99
Scan	81	96	72	99	99.99	99.99
Theft	97	94	90	99	99.99	99.99

TABLE V. F SCORE FOR SUB-CATEGORY CLASSIFICATION

	GaussianNB	LDA	Logistic Regression	Decision Tree	Random Forest	Ensemble
Normal	85	97	94	99.99	99.99	99.99
DDoS-HTTP	44	51	50	99.99	99.99	99.99
DDoS-TCP	72	98	90	99.99	99.99	99.99
DDoS-UDP	52	66	81	99.99	99.99	99.99
DoS-HTTP	95	58	51	99.99	99.99	99.99
DoS-TCP	86	69	88	99.99	99.99	99.99
DoS-UDP	49	78	66	99.99	99.99	99.99
OS	53	73	31	99.99	93.00	98.00
Service	50	83	47	99.99	92.00	99.00
Data Exfiltration	69	35	39	98.00	93.00	98.00
Key Logging	72	80	40	98.00	93.00	99.00

D. Feature Selection

Feature selection plays a vital role in machine learning. Feature selection is a process of selecting appropriate and essential features to improve the prediction capability of a model. Let suppose that a dataset with "n" number input features, then the feature selection outcome "s" features where "s" < "n". The feature selection technique reduces overfitting, improves the prediction power of a model, reduces training and testing time of the machine learning algorithm, and also decreases the difficulty level of the model [24]. There are three feature selection techniques: wrapper technique, filter technique, and embedded technique. In this paper, we used a Recursive Feature Elimination (RFE) technique to select important features from our proposed IoT botnet dataset. RFE is sometimes more expensive to run with a large number of input features, so we reduced the number of features by removing correlated features before using RFE [24],[25]. We used 0.70 as the correlation coefficient to remove the correlated features and detached eight correlated features, as shown in Table VI.

After removing all correlated features from the IoT botnet dataset, we applied the RFE technique to select relevant features. RFE used a random forest algorithm as an estimator for feature ranking. RFE technique rank features based on their importance. Fig. 4 shows feature importance using RFE for the full feature's dataset. We used a 3, 5- and 10-fold cross-validation test to verify the subset of selected features and the overfitting of the RFE model. We used accuracy, precision, recall, and F score as scoring metrics to evaluate our proposed RFE feature section model. The RFE model outcome for accuracy, precision, recall, and F score for feature selection presented in Fig. 5. We evaluated the 10, 20, and 40 best features of the IoT botnet dataset using RFE based on accuracy, precision, recall, and F score. The results were compared with the full features dataset. From our analysis, it is concluded that the dataset of 20 to 40 features is considered an optimal set of features. The accuracy, precision, recall, and F score for 20 features, and full features dataset for category and subcategory are presented in Table VII and Table VIII.

TABLE VI. CORRELATED FEATURES

Total	Feature Name
8	Bwd_Seg_Size_Avg, Fwd_Seg_Size_Avg, PSH_Flag_Cnt, Subflow_Bwd_Byts, Subflow_Bwd_Pkts, Subflow_Fwd_Byts, Subflow_Fwd_Pkts, URG_Flag_Cnt

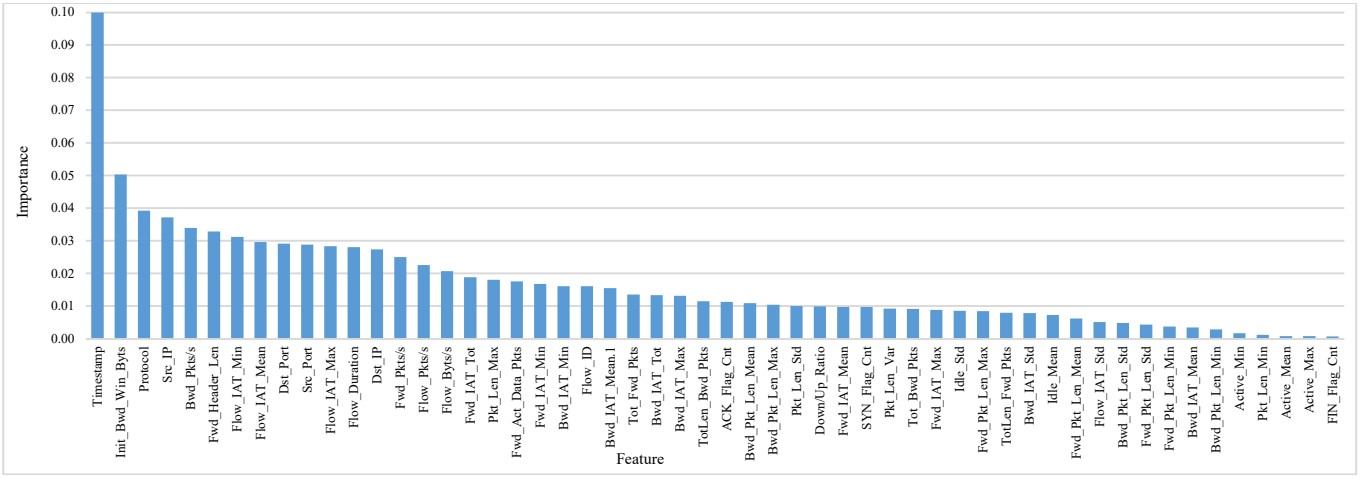

Fig.4. Feature Importance for All Features IoT Botnet Dataset

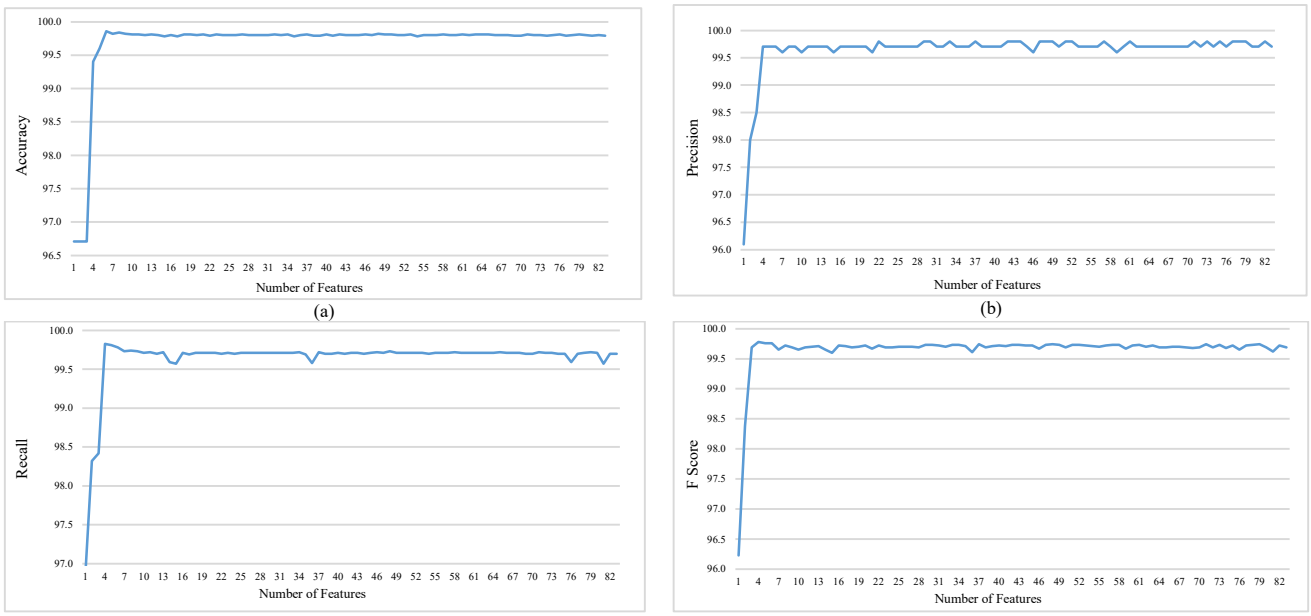

Fig.5. RFE (a) Accuracy, (b) Precision, (c) Recall and (d) F Score

TABLE VII. CATEGORY LABEL IOT BOTNET DATASET

	GaussianNB		LDA		Logistic Regression		Decision Tree		Random Forest		Ensemble	
	20 Features	*Full Features*	*20 Features*	*Full Features*	*20 Features*	*Full Features*	*20 Features*	*Full Features*	*20 Features*	*Full Features*	*20 Features*	*Full Features*
Accuracy	88	88	96	96	92	92	99.99	99.99	99.99	99.99	99.99	99.99
Precision	58	58	87	87	73	78	99.00	99.99	99.99	99.99	99.99	99.99
Recall	70	70	87	87	78	72	99.00	99.99	99.99	99.99	99.99	99.99
F1-Score	53	53	87	87	72	72	99.00	99.99	99.99	99.99	99.99	99.99

TABLE VIII. SUB-CATEGORY LABEL IOT BOTNET DATASET

	GaussianNB		LDA		Logistic Regression		Decision Tree		Random Forest		Ensemble	
	20 Features	*Full Features*	*20 Features*	*Full Features*	*20 Features*	*Full Features*	*20 Features*	*Full Features*	*20 Features*	*Full Features*	*20 Features*	*Full Features*
Accuracy	95	95	97	97	96	96	99.99	99.99	99.99	99.99	99.99	99.99
Precision	58	98	78	78	67	67	99.99	99.99	99.00	99.00	99.99	99.99
Recall	64	65	78	78	65	65	99.99	99.99	99.00	99.00	99.99	99.99
F1-Score	49	49	77	77	62	62	99.99	99.99	99.00	99.00	99.99	99.99

V. CONCLUSION

In this paper, we proposed a new botnet dataset for intrusion detection in IoT networks. Our proposed IoT botnet dataset generates new flow-based features and extra statistical features. IoT botnet dataset flow-based features provide a new foundation to develop, analyze, and evaluate a flow-based IDS effectively and efficiently for IoT networks. The IoT botnet dataset provides high ranking features for IoT networks to evaluate the malicious activity prediction models. Through the statistical analysis, three subsets of the IoT botnet dataset were generated using a recursive feature elimination algorithm, with 10, 20, and 40 best features. The validity of the dataset was evaluated via accuracy, precision, recall, and F score metrics. Our proposed IoT botnet dataset provides a reference point for detecting anomalous activity across IoT networks.

In the future, we plan to develop and evaluate a deep learning-based anomalous activity detection model using the IoT botnet dataset for intrusion detection in IoT networks.

REFERENCES

[1] C. Paquet, "Implementing Cisco IOS Network Security, " Cisco Press, 2012. http://www.ciscopress.com/store/implementing-cisco-ios-network-security-iins-640-554-9781587142727 (accessed Dec. 11, 2019).

[2] R. Luh, S. Marschalek, M. Kaiser, H. Janicke, and S. Schrittwieser, "Semantics-aware detection of targeted attacks: a survey," Journal of Computer Virology and Hacking Techniques, vol. 13, no. 1, pp. 47–85, 2017, doi: 10.1007/s11416-016-0273-3.

[3] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," IEEE Symposium on computational intelligence for security and defense applications, no. Cisda, pp. 1–6, 2009, doi: 10.1109/CISDA.2009.5356528.

[4] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," Computers & Security, vol. 86, pp. 147–167, 2019, doi: 10.1016/j.cose.2019.06.005.

[5] Y. Mirsky, T., Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection," pp. 18–21, 2018, doi: 10.14722/ndss.2018.23204.

[6] E. Alomari, S. Manickam, B. B. Gupta, P. Singh, and M. Anbar, "Design, deployment, and use of HTTP-based botnet (HBB) testbed," In 16th International Conference on Advanced Communication Technology, pp. 1265–1269, 2014, doi: 10.1109/ICACT.2014.6779162.

[7] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," Future Generation Computer Systems, vol. 100, pp. 779–796, 2019, doi: 10.1016/j.future.2019.05.041.

[8] I. Ullah and Q. H. Mahmoud, "A Scheme for Generating a Dataset for Anomalous Activity Detection in IoT Networks," in Goutte C., Zhu X. (eds) Advances in Artificial Intelligence. Canadian AI, 2020. Lecture Notes in Computer Science, vol 12109. Springer, Cham., 2020, pp. 508–520, doi: https://doi.org/10.1007/978-3-030-47358-7_52.

[9] S. Saad et al., "Detecting P2P botnets through network behavior analysis and machine learning," 2011 Ninth annual international conference on privacy, security, and trust., pp. 174–180, 2011, doi: 10.1109/PST.2011.5971980.

[10] I. Sharafaldin, A. H. Lashkari and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," ICISSP 2018 - 4th International Conference Information Systems Security and Privacy, vol. 2018-January, no. Cic, pp. 108–116, 2018, doi: 10.5220/0006639801080116.

[11] G. Thamilarasu and S. Chawla, "Towards deep-learning-driven intrusion detection for the internet of things," Sensors (Switzerland), vol. 19, no. 9, 2019, doi: 10.3390/s19091977.

[12] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy," 2019 International Carnahan Conference on Security Technology (ICCST), vol. 2019-October, no. Cic, 2019, doi: 10.1109/CCST.2019.8888419.

[13] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," Computer & Security, vol. 31, no. 3, pp. 357–374, 2012, doi: 10.1016/j.cose.2011.12.012.

[14] Y. Meidan et al., "N-BaIoT-Network-based detection of IoT botnet attacks using deep autoencoders," IEEE Pervasive Computing, vol. 17, no. 3, pp. 12–22, 2018, doi: 10.1109/MPRV.2018.03367731.

[15] L. Johansson and O. Olsson, "Improving Intrusion Detection for IoT Networks," MS thesis. 2018.

[16] I. Ullah and Q. H. Mahmoud, "An intrusion detection framework for the smart grid," 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 1–5, 2017, doi: 10.1109/CCECE.2017.7946654.

[17] I. Ullah and Q. H. Mahmoud, "A Two-Level Hybrid Model for Anomalous Activity Detection in IoT Networks," 16th IEEE Annual Consumer Communications & Networking Conference (CCNC), pp. 1–6, 2019, doi: 10.1109/CCNC.2019.8651782.

[18] I. Ullah and Q. H. Mahmoud, "A two-level Flow-based Anomalous Activity Detection System for IoT Networks," Electronics, vol. 9, no. 3, 2020, doi: 10.3390/electronics9030530.

[19] H. Chauhan, V. Kumar, S. Pundir, and E. S. Pilli, "A comparative study of classification techniques for intrusion detection," 2013 International Symposium on Computational and Business Intelligence, pp. 40–43, 2013, doi: 10.1109/ISCBI.2013.16.

[20] S. Duque and M. N. Bin Omar, "Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)," Procedia Computer Science, vol. 61, pp. 46–51, 2015, doi: 10.1016/j.procs.2015.09.145.

[21] W. Lee and S. J. Stolfo, "A Framework for Constructing Features and Models for Intrusion Detection Systems, " ACM transactions on Information and system security (TiSSEC), vol. 3, no. 4. 2000.

[22] I. Ullah and Q. H. Mahmoud, "IoT-Botnet Dataset," 2020. https://sites.google.com/view/iotbotnetdataset.

[23] A. H. Lashkari, G. D. Gil, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of tor traffic using time based features," ICISSP 2017-3rd International Conference Information Systems Security and Privacy,, vol. 2017-February, no. Cic, pp. 253–262, 2017, doi: 10.5220/0006105602530262.

[24] I. Ullah and Q. H. Mahmoud, "A filter-based feature selection model for anomaly-based intrusion detection systems," In 2017 IEEE International Conference on Big Data (Big Data), vol. 2018-Janua, pp. 2151–2159, 2017, doi: 10.1109/BigData.2017.8258163.

[25] I. Ullah and Q. H. Mahmoud, "A Hybrid Model for Anomaly-Based Intrusion Detection System," 2017 IEEE International Conference on Big Data (Big Data), vol. 2018-Janua, pp. 2160–2167, 2017, doi: 10.1109/BigData.2017.8258164.