Detect-IoT: A Comparative Analysis of Machine Learning Algorithms for Detecting Compromised IoT Devices

Yuba R. Siwakoti and Danda B. Rawat
Department of Electrical Engineering and Computer Science, Howard University
Washington, DC, 20059, USA
yuba.siwakoti@bison.howard.com,danda.rawat@howard.edu

ABSTRACT

The rapid expansion of IoT brings unmatched convenience and connectivity, but it also raises significant security concerns. The prioritization of functionality over security in IoT devices exposes vulnerabilities like default credentials, outdated components, and insecure interfaces. To mitigate risks and combat cyberattacks effectively, it is crucial to identify and isolate compromised IoT infrastructures. In this paper, we present a curated dataset for IoT security research, which combines 40 recent IoT behavior datasets using class balancing and feature reduction techniques. This curated dataset serves as a valuable resource for future research in the field. Additionally, we compare machine learning techniques to detect compromised IoT devices, leveraging preprocessed and SMOTE-balanced network data. Our ensemble model surpasses other methods, achieving an impressive up to 98 percent F1-score, thus highlighting its efficacy in predicting compromised IoT devices and emphasizing the significance of our dataset and methodology contributions.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Security and privacy**;

KEYWORDS

Security and Privacy, IoT security, Detect Compromised IoT Infrastructure, Computing Methodologies, Machine Learning, Network Behavioral Data, Enhanced IoT data

ACM Reference Format:
Yuba R. Siwakoti and Danda B. Rawat. 2023. Detect-IoT: A Comparative Analysis of Machine Learning Algorithms for Detecting Compromised IoT Devices. In *The Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '23), October 23–26, 2023, Washington, DC, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3565287.3616529

MobiHoc '23, October 23–26, 2023, Washington, DC, USA

ACM ISBN 978-1-4503-9926-5/23/10…$15.00
https://doi.org/10.1145/3565287.3616529

1 INTRODUCTION

According to estimates, millions of IoT devices remain susceptible to compromise due to programming practices and the use of outdated components [1]. The primary security risks stem from weak credentials and exposure to unnecessary services [11].

Over 80 percent of businesses and organizations utilize IoT to address various challenges, with nearly 20 percent having already encountered IoT-related attacks [14]. Cybercriminals exploit diverse attack vectors such as phishing, social engineering, man-in-the-middle (MitM) attacks, ransomware, and malware to compromise a significant number of IoT devices for malicious purposes (e.g., Dyn 2016, The Verkada Hack 2021, Stuxnet, etc.). Shodan exposes millions of IoT devices, among which thousands are compromised and abused for credential stuffing attacks and financial crimes [15]. Detecting and isolating vulnerable, exploitable, and compromised IoT infrastructures is imperative to secure network communications.

Various methodologies have been employed to detect vulnerable, exploitable, and compromised IoT devices. The primary objective of this study is to create effective ML/DL models to detect compromised IoT infrastructures.

Our Contributions:

(1) We used a recent dataset on Vulnerability and Attack Repository for the Internet of Things (VARIoT) [7, 10] that has not been used before to detect compromised IoT devices.
(2) We present curated enhanced data sourced from a network behavioral dataset, preserving performance by removing two-thirds of less-significant features for IoT devices under normal and compromised conditions.
(3) We employ Machine Learning (ML), Deep Learning (DL), and ensemble models to detect compromised IoT infrastructures using network behavioral data.

Paper Organization: The rest of the paper is organized as follows: Section 2 presents related literature followed by research methodology in section 3. Section 4 demonstrates and analyzes the results of this study followed by discussion in section 5. Limitations of this research and future prospects are discussed in 6. Finally, section 7 concludes the paper.

2 RELATED WORK

The exploit prediction model combines fastText and lightGBM algorithms, inspired by the state-of-the-art exploit prediction model [9]. Additionally, a novel approach utilizing a natural language model was proposed in [16], extracting discussions from the dark web/deep web to predict vulnerability exploitation. ML-based phishing attack detection was proposed in [4].

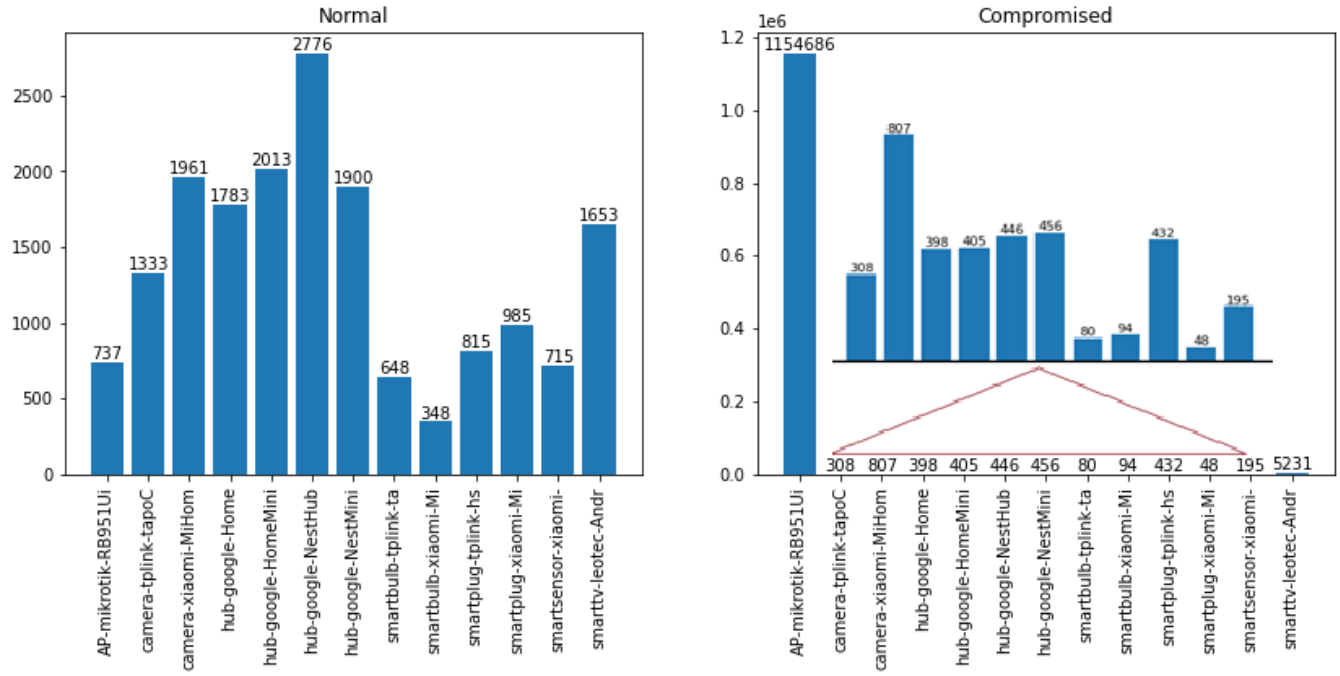

Figure 1: Distribution of original imbalanced data

Moustafa et al. [12] introduced an automatic method for identifying vulnerabilities and exploitation through network flow analysis. They designed an Industrial IoT (IIoT) testbed with IoT gateways. Ullah et al. [17] also utilized ML techniques to identify IoT devices. Al-Boghdady et al. [3] developed a random forest-based ML system for detecting vulnerabilities in the C/C++ code of IoT devices' operating systems. Duan et al. [8] proposed an automated security assessment framework for IoT, leveraging natural language processing and ML to predict vulnerability metrics.

A logistic regression-based solution was proposed to identify compromised IoT devices within a botnet [13]. Da-cruz et al. [6] presented a solution to detect compromised IoT devices through attacks that obtain device credentials, such as replication attacks, by analyzing abnormal network traffic.

None of the mentioned studies utilized the recent IoT dataset (VARIoT) [7, 10] consisting of compromised and normal network behavior traffic. Additionally, we conducted a comparative analysis of various models, assessing performance changes with different feature sets, and presenting a feature-reduced enhanced dataset.

3 RESEARCH METHODOLOGY

In this section, we will provide a brief overview of the dataset used in this paper, followed by an outline of our methodology.

Dataset: In this study, we utilize the recent VARIoT [7], encompassing network traffic data for IoT devices under both normal and compromised conditions.

The VARIoT dataset is generated by the data extraction laboratory [2], as part of the VARIoT project co-financed by the Connecting Europe Facility of the European Union. We select 40 datasets for IoT devices that include entries for both normal and compromised conditions, covering a diverse range of thirteen devices from seven distinct IoT categories. The encompassed IoT device categories comprise smart cameras, smart TVs, smart Hubs, smart plugs, smart bulbs, smart sensors, and access points (AP).

The selected dataset comprises 1.18 million entries, consisting of 115 features including device_type and label. Specifically, the dataset consists of 1.16 million entries classified as compromised and 17.6 thousand entries categorized as normal, as illustrated in Figure 1.

Methodology: The methodology proposed in this paper is illustrated in Figure 2. Initially, we download data from VARIoT [7], which is significantly imbalanced and dominated by compromised class. To address this issue, we apply both under-sampling and Synthetic Minority Over-sampling Technique (SMOTE) [5]. Specifically, out of the 1.16 million compromised entries, a vast majority (1.15 million) belong to a single IoT device (AP-Microtik), which we under-sample to approximately 2000 entries (about average data point for other devices), as indicated in the initial data distribution in Figure 1. After this under-sampling step, the dataset consists of 17.6 thousand normal data points and 10.6 thousand compromised data points. This processed data is referred to as the partially balanced dataset.

In our pre-processing phase, we execute several steps, including removing empty or irrelevant features, handling missing and outlier values, and normalizing the data. After eliminating features with the majority of instances as empty/NaN, the preprocessed dataset consists of 53 features including the label. Subsequently, we apply SMOTE oversampling to both the normal and compromised

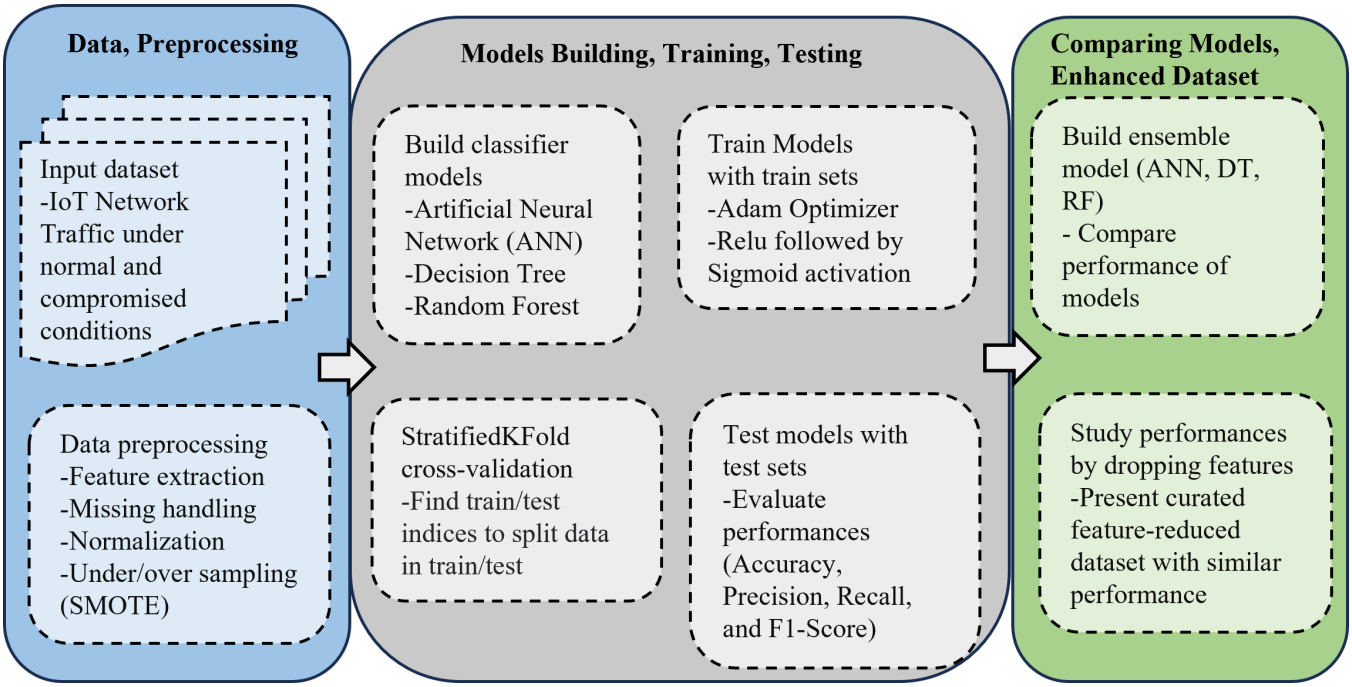

Figure 2: Proposed Methodology

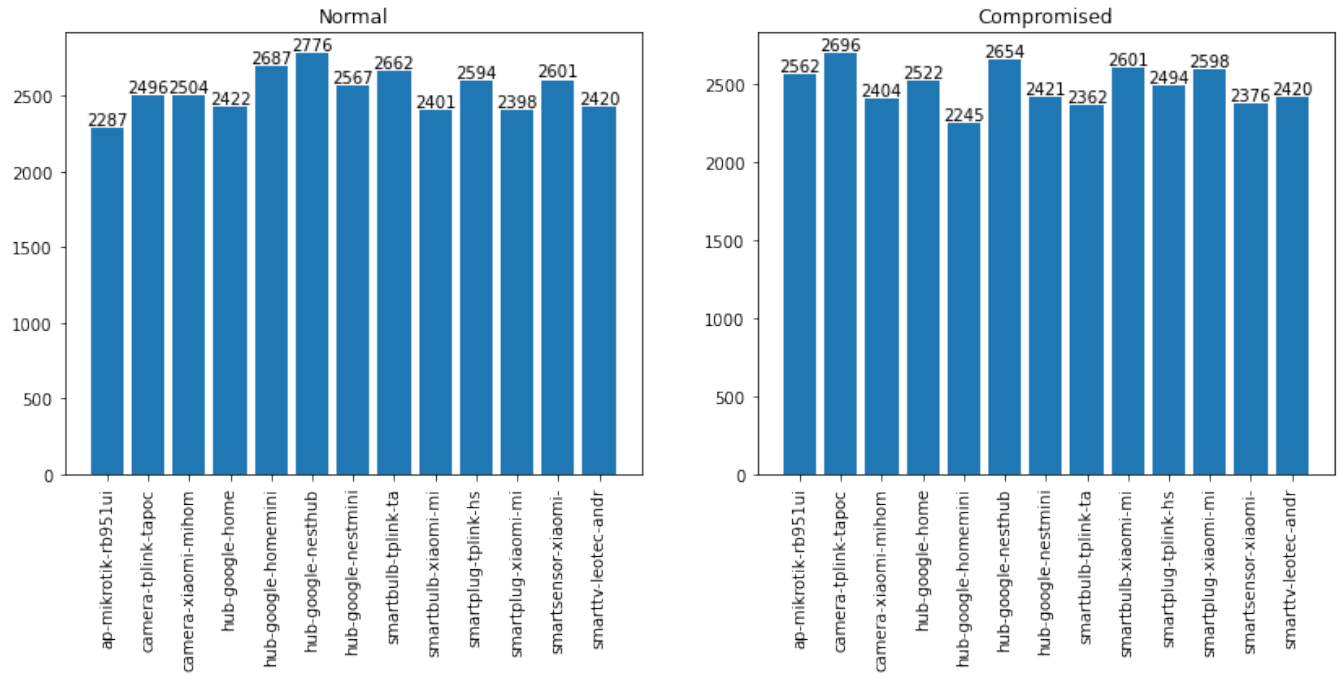

Figure 3: Visualization of the balanced data distribution after SMOTE

datasets, resulting in approximately 32,000 instances for each class as depicted in Figure 3.

To further prepare the data, we encode categorical data using LabelEncoder and handle missing values in certain rows by substituting them with the median. We then utilize MinMaxScaler and

StandardScaler to normalize data for categorical and numerical features, respectively.

We constructed three classifiers: Artificial Neural Network (ANN), Decision Tree (DT), and Random Forest (RF). To optimize model performance, we utilized Stratified 5-fold cross-validation to identify the best train/test indices for data splitting. Subsequently, we trained all three models using the designated train sets and implemented early stopping rules. For the ANN, we employed the RELU activation function in the hidden layers, Sigmoid in the output layer, and Adam as the optimizer to enhance efficiency.

As shown in Figure 2, we assess all three models using four performance metrics - accuracy, precision, recall, and F1-score - by applying the models to predict outcomes in the test sets. Subsequently, we construct an ensemble model incorporating all ANN, DT, and RF classifiers, and evaluate its performance by aggregating predictions from the three models using the majority voting approach.

Lastly, we examine the impact of feature reduction on all models to assess the significance of individual features in their performance. We systematically drop features one by one and observe changes in F1-score and loss, investigating the dependence and independence of features with respect to the models.

4 RESULTS AND ANALYSIS

As outlined in Section 3, we conduct a performance comparison of four ML/DL models: DT, ANN, RF, and an ensemble of these three models. The evaluation encompasses accuracy, precision, recall, and F1-score metrics across three datasets: the unbalanced dataset, the partially balanced dataset (under-sampled), and the (SMOTE) balanced dataset, as illustrated in Figure 4. The performance metrics accuracy, precision, recall, and F1-score, are computed using equations 1, 2, 3, and 4, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

While accuracy is a commonly used metric for evaluating model performance, it can be misleading, especially in class-biased data. For instance, in the unbalanced dataset, all classifiers may exhibit high accuracy above 99 percent, but their recall and precision remain low (about 95 percent) which is improved slightly in the partially_balanced dataset (by 1 percent) and more in the SMOTE_balanced dataset (by 3 percent), as illustrated in Fig 4. We primarily focus on the F1-score metric in our analysis due to its consideration of both precision and recall (as shown in equation 4). Moreover, the ensemble method, which considers the majority of votes from all models, provides a more rational and consistent performance.

Across all four models, all performance metrics, including accuracy, precision, recall, and F1-score, exceed 94 percent, reflecting quality data instances and associated preprocessing and model selection. However, in the unbalanced dataset, the performance parameters display inconsistency. On the other hand, in the balanced dataset, these parameters are comparatively higher and more consistent. Notably, in the SMOTE_balanced dataset, the F1-score surpasses 97 percent for all models, and the ensemble method outperforms the others with an impressive F1-score of over 98 percent.

Furthermore, we conduct an analysis to understand the impact of feature reduction on the F1-score and loss (MSE: Mean Squared Error) for the SMOTE balanced dataset. To identify significant features for models' performance, we systematically remove one feature at each step and measure both the loss and F1-score, as presented in Figure 5.

In the figure, we observe minimal changes in both the F1-score and loss until we drop 35 less significant features. Beyond this point, the F1-score starts to decrease, while the loss increases sharply with a slight variation in rate across different models. To validate our findings, we also examine the correlation of features with the label. This analysis confirms that the 35 removed features exhibit a very low correlation with the label, thereby contributing little to the model's overall performance.

5 DISCUSSION

For our study, we opted for a relatively new and less explored dataset [7], which was generated by the data extraction laboratory [2] as part of the VARIoT project, co-financed by the Connecting Europe Facility of the European Union. This dataset served as the basis for our pre-processing steps to prepare it for ML/DL models.

Following an exploratory analysis of the dataset, we diligently cleaned it and performed feature engineering to select the essential features for our model. The pre-processing phase meticulously addressed various issues, such as handling empty features, managing missing, duplicate, and outlier data points, addressing data/class imbalances, and carrying out encoding and normalization.

Once the pre-processing was completed, we developed several models (ANN, DT, RF, and ensemble) to detect compromised IoT infrastructure. To optimize the training and testing process, we selected the best indices using stratified 5-fold cross-validation. The models were then trained using the selected indices.

The performance of these models on the pre-processed and balanced dataset displayed remarkable results, as depicted in Fig 4. The efforts put into data cleaning and feature engineering significantly contributed to the models' overall effectiveness in predicting compromised IoT infrastructure.

Despite achieving commendable performances for all models, the ensemble model stands out as the top performer, attaining an impressive F1-score exceeding 98 percent on a (SMOTE) balanced dataset. Moreover, the ensemble model maintains a remarkable precision level of approximately 99 percent in the balanced dataset. The ensemble approach proves to be highly effective in predicting compromised IoT infrastructure, showcasing its superiority over individual models in this context.

Furthermore, we delved into the significance of features in determining the models' performances and identified that only 17 features hold substantial relevance for the models. This observation was validated by analyzing feature correlations with the label. Surprisingly, dropping 35 less relevant features did not significantly impact the models' performance, indicating that we can achieve

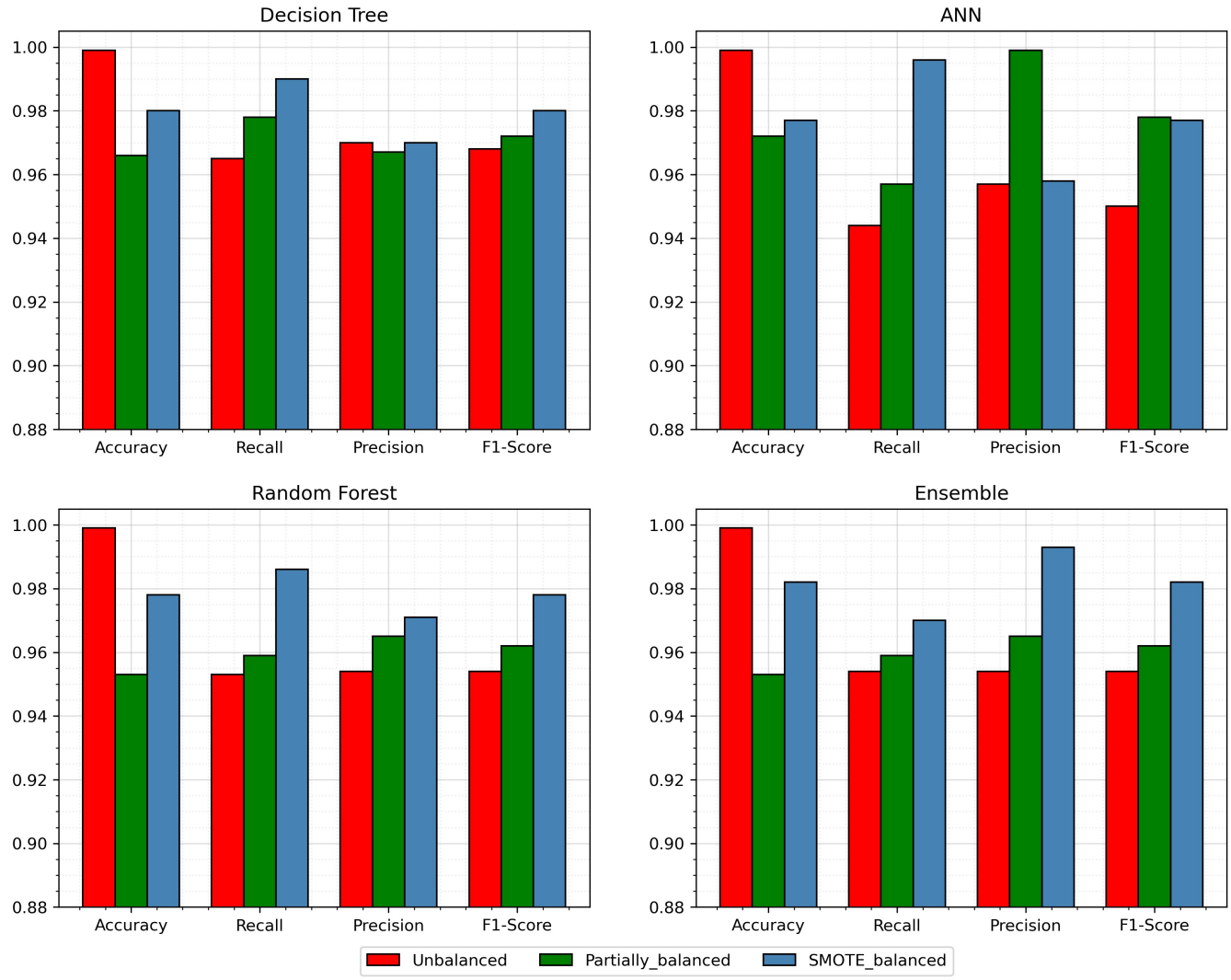

Figure 4: Performance comparison of ML/DL models to detect compromised IoT devices

excellent results by focusing on only these 17 key features. This insight provides an opportunity to reduce cost and time while collecting Netflow data with only essential features instead of the original 113 features used in the VARIoT [7] dataset.

Building upon the insights gained from our observations, we introduced the cleaned and feature-engineered dataset[1] to the research community. This valuable resource comprises a total of 18 features, inclusive of the label, and holds immense potential for further exploration and research.

6 LIMITATIONS AND FUTURE PROSPECTS

Limitations: Our dataset comprises only 13 devices representing 7 categories of IoT. However, in the vast IoT ecosystem, there exist millions of IoT devices across thousands of IoT categories. Generating data for all these devices is infeasible due to time and resource complexity.

Validating the model's performance using additional real and synthetic datasets is crucial to enhance its reliability and usability. However, for this study, we lacked such validation.

Future Prospects: Given the aforementioned limitations, a potential approach to address them is by exploring synthetic data generation techniques that closely resemble the data used in this paper. Additionally, for future data generation, a well-defined sampling strategy can be employed to ensure representative IoT device selection, mitigating potential biases and improving the dataset's overall quality.

[1]Enhanced dataset will be shared in GitHub once the paper will go public.

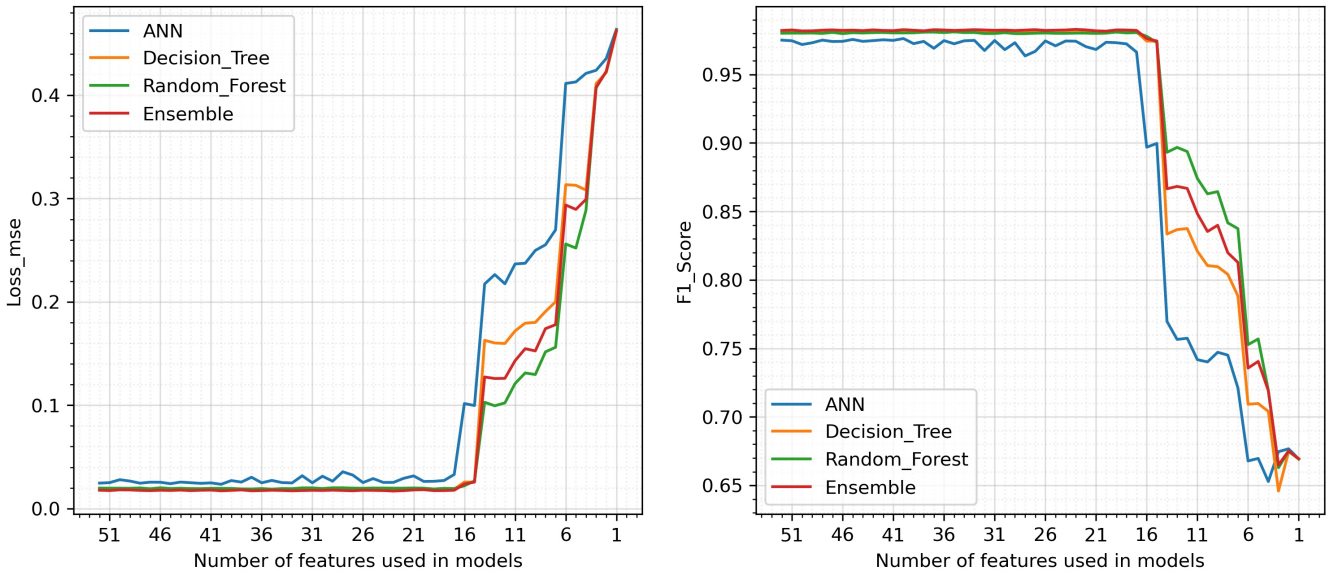

Figure 5: Analyzing F1-score and loss variations upon feature reduction in SMOTE balanced dataset

7 CONCLUSION

In this paper, we apply ML models such as ANN, RF, DT, and ensemble to detect compromised IoT devices using a Network behavioral dataset. To address the class-imbalanced dataset, we employed under-sampling and SMOTE. Our ensemble model outperforms others with over 98 percent F1-score while maintaining consistency.

Considering the significance of features on model performance, we identify 17 significant features and introduce a curated dataset with only those features and the label encompassing the network behavior of 13 IoT devices under both normal and compromised conditions. This dataset will be available to the research community that facilitates further investigations in the realm of IoT security.

ACKNOWLEDGEMENT

This work was supported in part by the U.S. National Science Foundation under Grant CNS/SaTC 2039583, and in part by the DoD Center of Excellence in AI and Machine Learning (CoE-AIML) at Howard University under Contract W911NF-20-2-0277 with the U.S. Army Research Laboratory and MasterCard Research Funds. However, any opinion, finding, conclusions, or recommendations expressed in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the funding agencies.

REFERENCES

[1] 2020. Amnesia:33 Identify and Mitigate the Risk From Vulnerabilities Lurking in Millions of IoT, OT and IT Device. https://www.forescout.com/research-labs/amnesia33/

[2] 2021. *Data extraction laboratory – VARIoT.* Retrieved May 10, 2023 from https://www.variot.eu/2021/10/07/data-extraction-laboratory/

[3] Abdullah Al-Boghdady, Mohammad El-Ramly, and Khaled Wassif. 2022. iDetect for vulnerability detection in internet of things operating systems using machine learning. *Scientific Reports* 12, 1 (2022), 17086.

[4] Manish Bhurtel, Yuba R Siwakoti, and Danda B Rawat. 2022. Phishing Attack Detection with ML-Based Siamese Empowered ORB Logo Recognition and IP Mapper. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 1–6.

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[6] Mauro AA da Cruz, Lucas R Abbade, Pascal Lorenz, Samuel B Mafra, and Joel JPC Rodrigues. 2022. Detecting Compromised IoT Devices Through XGBoost. *IEEE Transactions on Intelligent Transportation Systems* (2022).

[7] Universidad de Mondragón. 2021. *IoT security - network traffic under normal and compromised conditions, Dataset.* Retrieved May 10, 2023 from https://data.europa.eu/data/datasets?keywords=variot&locale=en

[8] Xuanyu Duan, Mengmeng Ge, Triet Huynh Minh Le, Faheem Ullah, Shang Gao, Xuequan Lu, and M Ali Babar. 2021. Automated security assessment for the internet of things. In *2021 IEEE 26th Pacific Rim International Symposium on Dependable Computing (PRDC)*. IEEE, 47–56.

[9] Yong Fang, Yongcheng Liu, Cheng Huang, and Liang Liu. 2020. FastEmbed: Predicting vulnerability exploitation possibility based on ensemble machine learning algorithm. *Plos one* 15, 2 (2020), e0228439.

[10] Marek Janiszewski, Marcin Rytel, Piotr Lewandowski, and Hubert Romanowski. 2022. VARIoT - Vulnerability and Attack Repository for the Internet of Things. In *2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*. 752–755. https://doi.org/10.1109/CCGrid54584.2022.00085

[11] Deepak Kumar, Kelly Shen, Benton Case, Deepali Garg, Galina Alperovich, Dmitry Kuznetsov, Rajarshi Gupta, and Zakir Durumeric. 2019. All things considered: an analysis of IoT devices on home networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 1169–1185.

[12] Nour Moustafa, Benjamin Turnbull, and Kim-Kwang Raymond Choo. 2018. Towards automation of vulnerability and exploitation identification in IIoT networks. In *2018 IEEE International Conference on Industrial Internet (ICII)*. IEEE, 139–145.

[13] Anton O Prokofiev, Yulia S Smirnova, and Vasiliy A Surov. 2018. A method to detect Internet of Things botnets. In *2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*. IEEE, 105–108.

[14] Gartner Research. 2020. *IoT Security Primer: Challenges and Emerging Practicess.* Technical Report.

[15] Yuba Raj Siwakoti, Manish Bhurtel, Danda B. Rawat, Adam Oest, and R. C. Johnson. 2023. Advances in IoT Security: Vulnerabilities, Enabled Criminal Services, Attacks, and Countermeasures. *IEEE Internet of Things Journal* 10, 13 (2023), 11224–11239. https://doi.org/10.1109/JIOT.2023.3252594

[16] Nazgol Tavabi, Palash Goyal, Mohammed Almukaynizi, Paulo Shakarian, and Kristina Lerman. 2018. Darkembed: Exploit prediction with neural language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[17] Imtiaz Ullah and Qusay H Mahmoud. 2021. Network traffic flow based machine learning technique for IoT device identification. In *2021 IEEE International Systems Conference (SysCon)*. IEEE, 1–8.