An Ensemble Approach for Detecting IoT Device Compromise Using Network Behavioral Data

Yuba R. Siwakoti, Danda B. Rawat

***Abstract*—The rapid proliferation of IoT devices has brought unparalleled convenience and connectivity to various domains. However, this exponential growth in IoT deployment has also raised serious security concerns. IoT devices are often designed with greater emphasis on functionality and performance, relegating security measures to a lower priority. Consequently, a host of vulnerabilities plague these devices, including default credentials, outdated components, lack of regular updates, insecure network interfaces, and exposed web/API ecosystem interfaces.**

These inherent weaknesses provide fertile ground for malicious actors to compromise a large number of IoT devices, leveraging them for nefarious purposes. The consequences are severe, as these compromised devices can serve as entry points for cyberattacks, posing significant threats to both individuals and organizations. It becomes crucial to isolate compromised IoT infrastructures in the wild to mitigate the potential risks and reduce attacks initiated from or associated with these vulnerable devices.

To address this pressing security concern, this paper introduces an innovative ensemble approach that harnesses the power of machine learning to predict compromised IoT infrastructures accurately. Leveraging curated network behavioral data, our proposed method presents the comparative performances of machine learning techniques to identify compromised IoT infrastructure. By employing a combination of machine learning techniques, our model achieves an impressive up to 98 percent F1-score, demonstrating its effectiveness in predicting compromised devices.

***Index Terms*—IoT security, Machine Learning, Detect Compromised IoT infrastructure, Vulnerable, Exploitable, Ensemble Approach, Network Behavioral Data**

I. 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to estimates, millions of IoT devices remain susceptible to compromise due to programming practices and the use of outdated components [1].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 [2]."

Over 80 percent of businesses and organizations utilize IoT to address various challenges, with nearly 20 percent having already encountered IoT-related attacks [3]. 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 [4]. 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 new dataset that has not been used before to detect compromised IoT devices.
2) We employ machine learning techniques (Decision Tree, Random Forest), deep learning (ANN), and ensemble models to detect compromised IoT infrastructures using network behavioral data.
3) 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.

Paper Organization: The rest of the paper is organized as follows: Section II presents related literature followed by methodology in section III. Section IV demonstrates and analyzes the results of this study followed by discussion in section V. Limitations of this research and future prospects are discussed in VI. Finally, section VII concludes the paper.

II. Related Work

The exploit prediction model combines fastText and lightGBM algorithms, inspired by the state-of-the-art exploit prediction model [5]. Additionally, a novel approach utilizing a natural language model was proposed in [6], extracting discussions from the dark web/deep web to predict vulnerability exploitation.

"Moustafa et al. [7] 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. [8] also utilized machine learning techniques to identify IoT devices. Al et al. [9] developed a Random Forest-based machine learning system for detecting vulnerabilities in C/C++ code of IoT devices' operating systems. Duan et al. [10] proposed an automated security assessment framework for IoT, leveraging natural language processing and machine learning to predict vulnerability metrics."

A logistic regression-based solution was proposed to identify compromised IoT devices within a botnet [11]. Da et al. [12] presented a solution to detect compromised IoT devices through attacks that obtain device credentials, such as replication attacks, by analyzing abnormal network traffic. Yavuz et al. [13] introduced a deep learning-based approach for detecting routing attacks in IoT and generated attack data using the Cooja IoT simulator for 10 to 1000 IoT nodes.

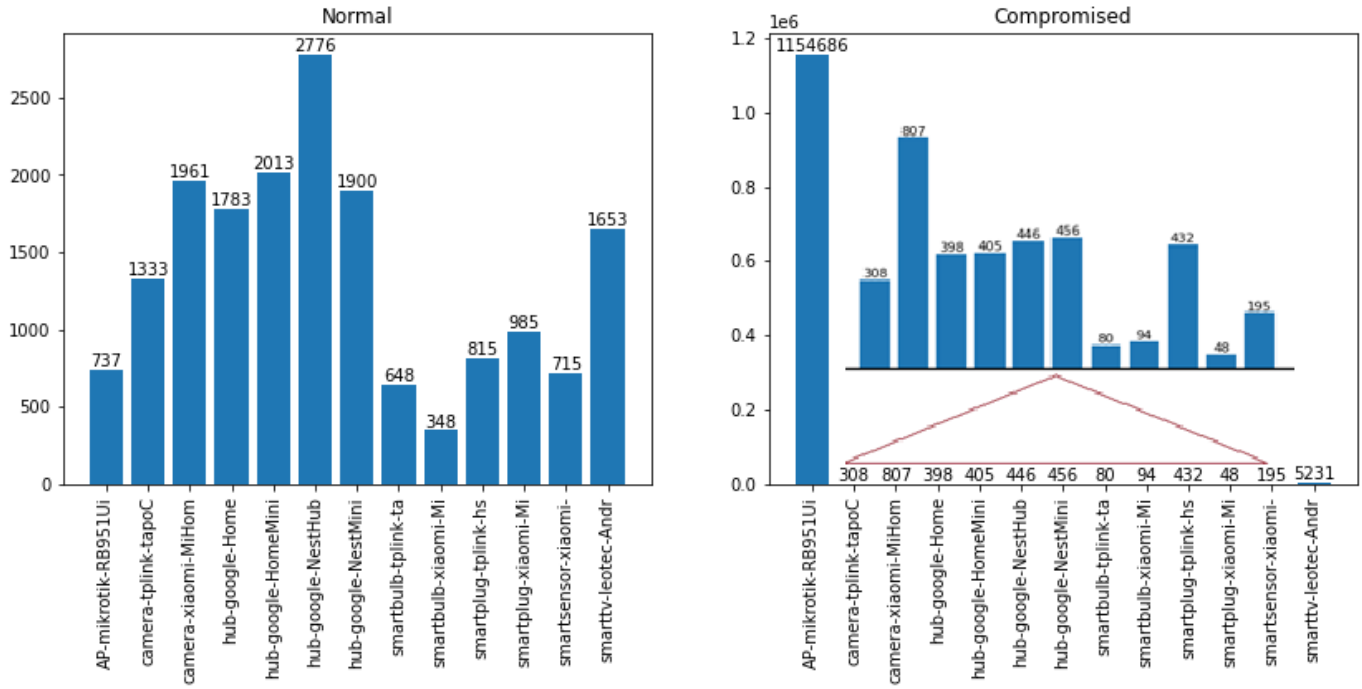

Fig. 1. Distribution of original imbalanced data

None of the mentioned studies utilized the VARIoT dataset, which comprises both compromised and normal flow-based traffic. Additionally, we conducted a comparative analysis of various models, assessing performance changes with different feature sets to determine feature importance for the models.

III. Proposed Approach

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 paper, we utilize the recent Vulnerability and Attack Repository for IoT (VARIoT) dataset [14], encompassing network traffic data for IoT devices under both normal and compromised conditions. The VARIoT dataset is generated by the data extraction laboratory [15], as part of the VARIoT project co-financed by the Connecting Europe Facility of the European Union. Among the 82 IoT network traffic behavior datasets (captured from TCPDUMP/LIBPCAP and extracted from Argus) that represent both legitimate and compromised conditions, we select 40 datasets. Each of these datasets contains entries for both normal and compromised conditions, encompassing thirteen devices from seven IoT categories, including smart camera, smart TV, smart Hub, smart plug, smart bulb, smart sensor, and AP.

The original 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 Fig. 1.

Methodology The methodology proposed in this paper is illustrated in Fig. 2. Initially, we download data from VARIoT [14], which is significantly imbalanced. To address this issue, we apply both under-sampling and Synthetic Minority Over-sampling Technique (SMOTE) [16]. 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 2500 entries (slightly higher than the data instances for the majority of selected IoT devices), as indicated in the initial data distribution in Fig. 1. After this under-sampling step, the dataset consists of 17.6 thousand normal data points and 11.4 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 NaN, the dataset consists of 53 features, along with the label. Subsequently, we apply SMOTE oversampling to both the normal and compromised datasets, resulting in approximately 32,000 instances for each class, with around 2500 entries for each device in a class, as depicted in Fig. 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

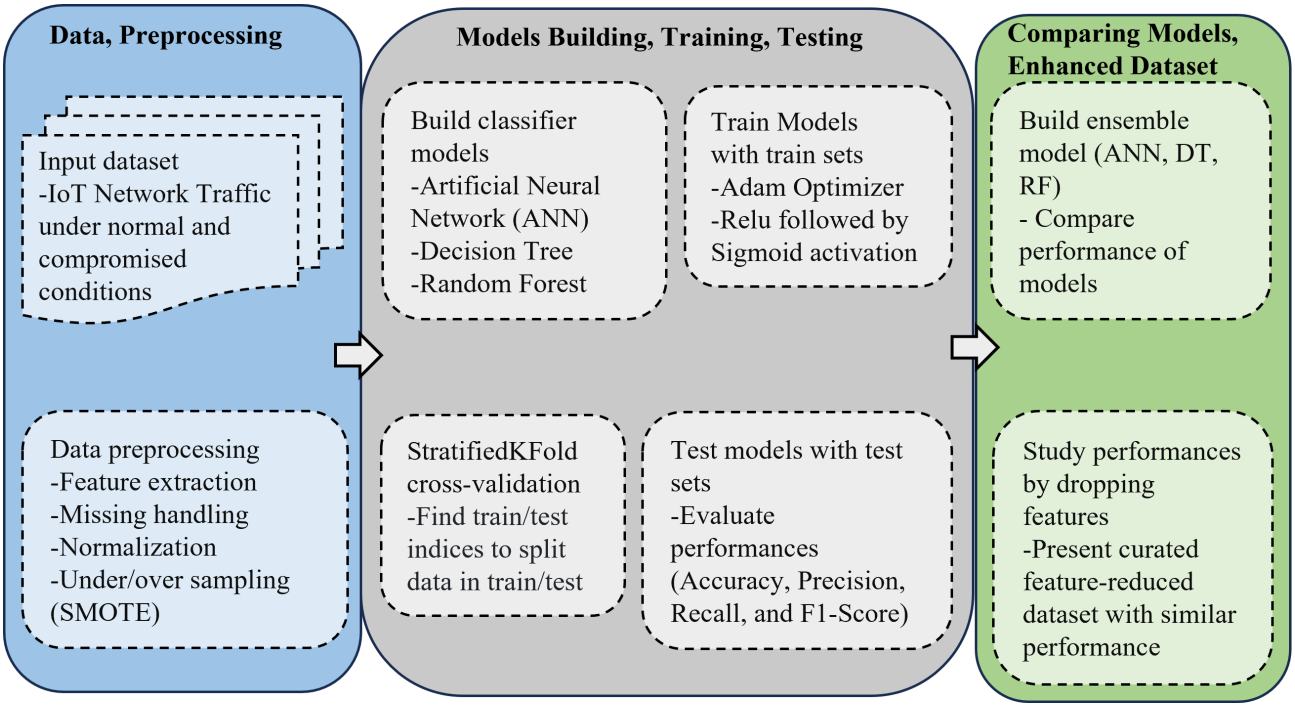

Fig. 2. Proposed Methodology

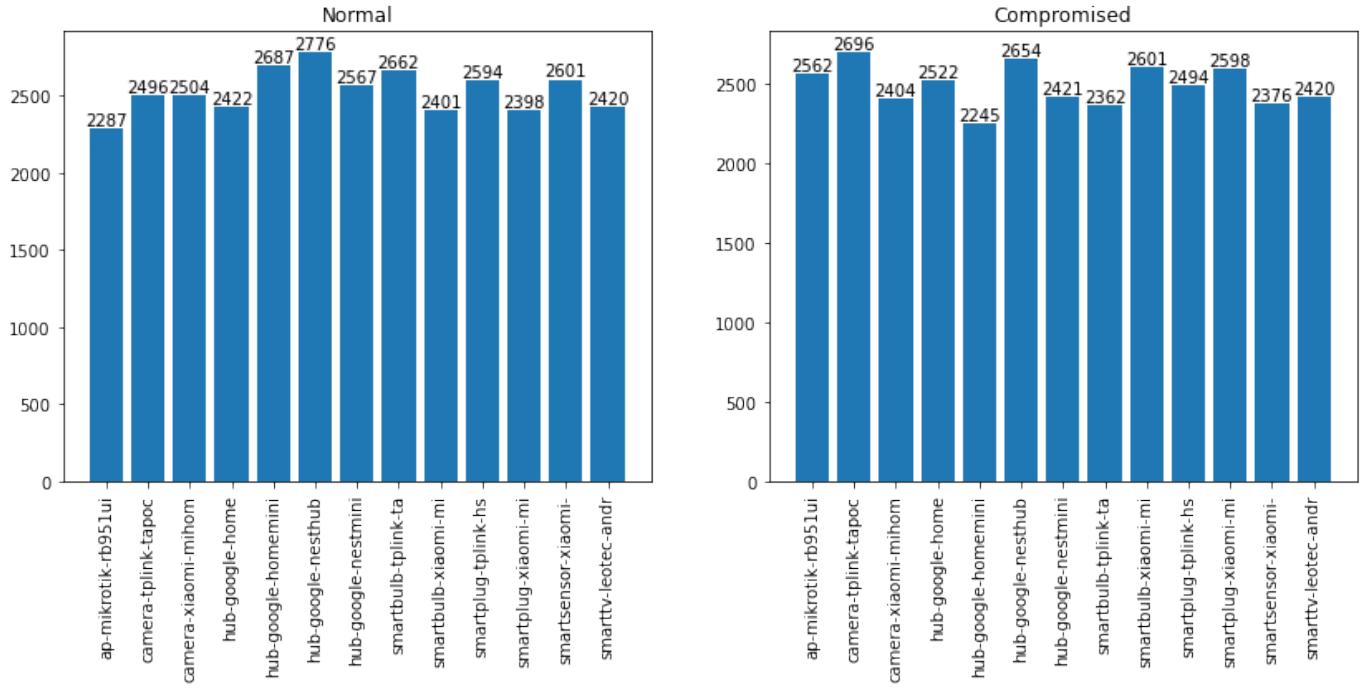

Fig. 3. Visualization of the balanced data distribution after SMOTE

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 Fig. 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.

IV. Results and Analysis

As outlined in Section III, we conduct a performance comparison of four ML/DL models: Decision Tree, Random Forest, Artificial Neural Network, 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 dataset (fully balanced), as illustrated in Fig. 4. The performance metrics accuracy, precision, recall, and F1-Score, are computed using equations 1, 2, 3, and 4, respectively.

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

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

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

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

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, but their recall and precision remain low. These metrics show a slight improvement in the partially_balanced dataset and a more significant improvement in the SMOTE_balanced dataset, as illustrated in Fig 4. Due to its consideration of both precision and recall (as shown in equation 4), we primarily focus on the F1-score in our analysis. Moreover, the ensemble method, which considers the majority of votes from all models, provides a more rational performance assessment.

Across all four models, all performance metrics, including accuracy, precision, recall, and F1-score, exceed 94 percent, reflecting excellent performance. 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 as anticipated, 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 (Mean Square Error) for the SMOTE balanced dataset. To identify significant features for predictions, we systematically remove one feature at each step and measure both the loss and F1-score, as presented in Fig. 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 sharply increases 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.

V. Discussion

For our study, we opted for a relatively new and less explored dataset [14], which was generated by the data extraction laboratory [15] 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 predict 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 fully balanced (SMOTE) 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 model performance, indicating that we can achieve excellent results by focusing on these 17 key features. This insight provides an opportunity to reduce cost and time by collecting Netflow data with only these

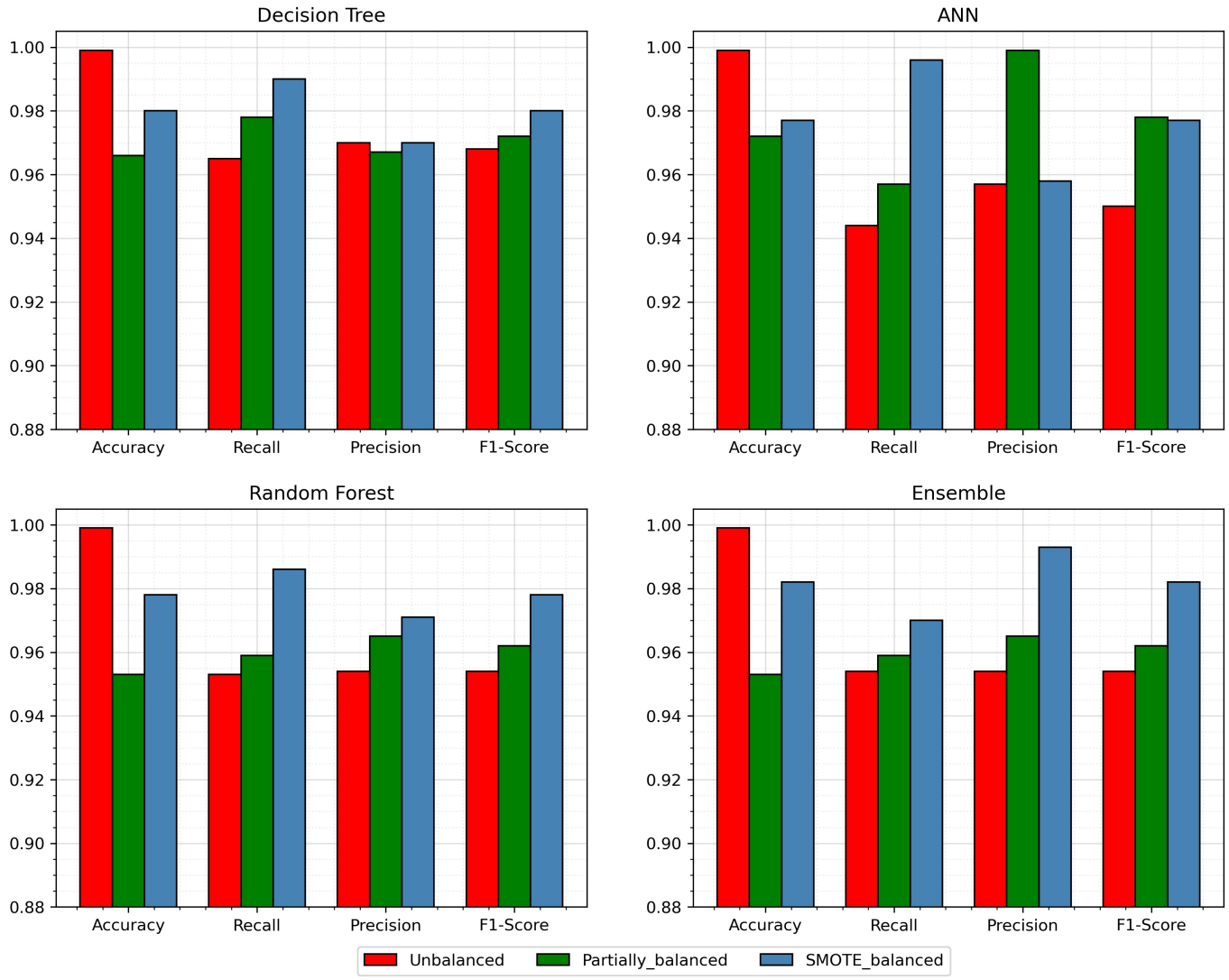

Fig. 4. Performance comparison of ML/DL models across all data sets

essential features instead of the original 113 features used in the VARIoT [14] dataset.

Building upon the insights gained from our observations, we proudly introduced the cleaned and feature-engineered dataset 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.

VI. Limitations and Future Prospects

Limitations: Our dataset comprises a limited number of 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. To ensure representativeness, it is preferable to include devices in the study following a well-defined sampling strategy, as discussed in [4].

To enhance the reliability and usability of the model, it is essential to validate its performance using both synthetic and real datasets. This dual validation approach will provide more robust and trustworthy results, further solidifying the model's efficacy in real-world scenarios. *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.

VII. Conclusion

In this paper, we apply ANN, RF, DT, and ensemble methods to detect compromised IoT devices using a Netflow behavioral dataset. To address the highly class-imbalanced dataset, we employed under-sampling and SMOTE for balancing. All models achieved F1-scores above 97 percent on

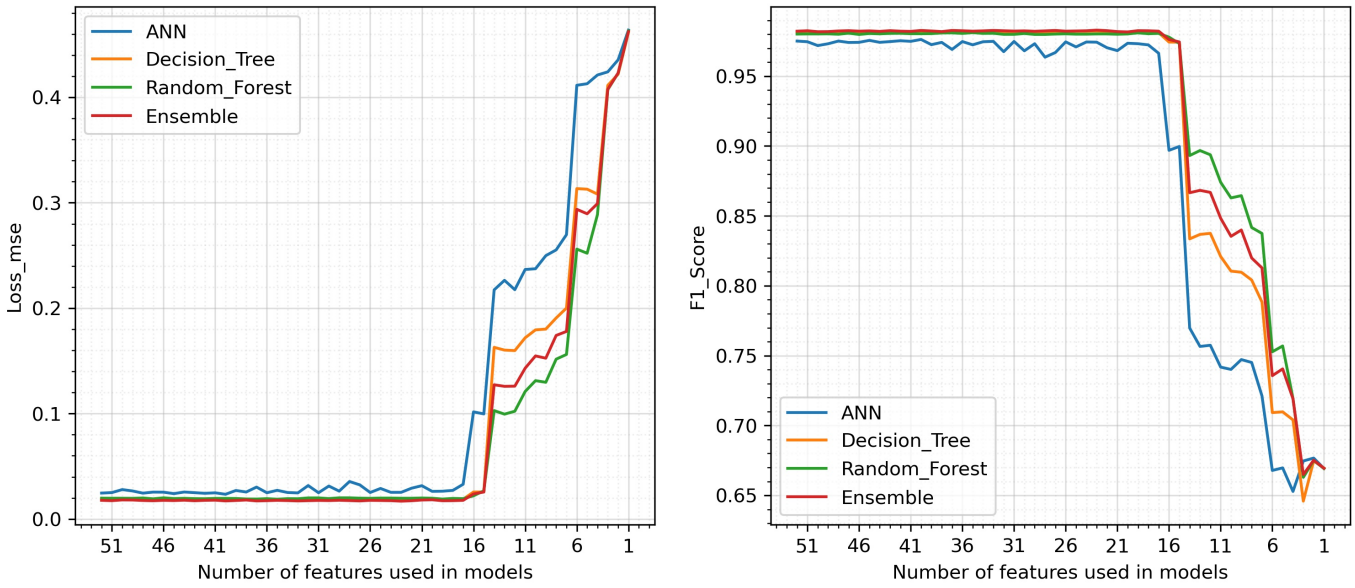

Fig. 5. Analyzing F1-score and loss variations upon feature reduction in SMOTE balanced dataset

the balanced dataset, with the ensemble outperforming at over 98 percent F1-score, and SMOTE enhancing consistency.

Considering the significance of features on model performance, we identified 35 less important and 17 influential features. Following preprocessing and feature engineering, we introduce a curated 18-feature dataset encompassing the network behavior of 13 IoT devices under both normal and compromised conditions. This dataset is made available to the research community, facilitating 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. 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] "Amnesia:33 Identify and Mitigate the Risk From Vulnerabilities Lurking in Millions of IoT, OT and IT Device," December 2020. [Online]. Available: https://www.forescout.com/research-labs/amnesia33/

[2] D. Kumar, K. Shen, B. Case, D. Garg, G. Alperovich, D. Kuznetsov, R. Gupta, and Z. Durumeric, "All things considered: an analysis of IoT devices on home networks," in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019, pp. 1169–1185.

[3] "Iot security primer: Challenges and emerging practices," https://www.gartner.com/en/doc/iot-security-primer-challenges-and-emerging-practices.

[4] Y. R. Siwakoti, M. Bhurtel, D. B. Rawat, A. Oest, and R. Johnson, "Advances in iot security: Vulnerabilities, enabled criminal services, attacks and countermeasures," *IEEE Internet of Things Journal*, pp. 1–1, 2023.

[5] Y. Fang, Y. Liu, C. Huang, and L. Liu, "Fastembed: Predicting vulnerability exploitation possibility based on ensemble machine learning algorithm," *Plos one*, vol. 15, no. 2, p. e0228439, 2020.

[6] N. Tavabi, P. Goyal, M. Almukaynizi, P. Shakarian, and K. Lerman, "Darkembed: Exploit prediction with neural language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[7] N. Moustafa, B. Turnbull, and K.-K. R. Choo, "Towards automation of vulnerability and exploitation identification in iiot networks," in *2018 IEEE International Conference on Industrial Internet (ICII)*. IEEE, 2018, pp. 139–145.

[8] I. Ullah and Q. H. Mahmoud, "Network traffic flow based machine learning technique for iot device identification," in *2021 IEEE International Systems Conference (SysCon)*. IEEE, 2021, pp. 1–8.

[9] A. Al-Boghdady, M. El-Ramly, and K. Wassif, "idetect for vulnerability detection in internet of things operating systems using machine learning," *Scientific Reports*, vol. 12, no. 1, p. 17086, 2022.

[10] X. Duan, M. Ge, T. H. M. Le, F. Ullah, S. Gao, X. Lu, and M. A. Babar, "Automated security assessment for the internet of things," in *2021 IEEE 26th Pacific Rim International Symposium on Dependable Computing (PRDC)*. IEEE, 2021, pp. 47–56.

[11] A. O. Prokofiev, Y. S. Smirnova, and V. A. Surov, "A method to detect internet of things botnets," in *2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*. IEEE, 2018, pp. 105–108.

[12] M. A. da Cruz, L. R. Abbade, P. Lorenz, S. B. Mafra, and J. J. Rodrigues, "Detecting compromised iot devices through xgboost," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[13] F. Y. Yavuz, Ü. Devrim, and G. Ensar, "Deep learning for detection of routing attacks in the internet of things," *International Journal of Computational Intelligence Systems*, vol. 12, no. 1, p. 39, 2018.

[14] "Variot – vulnerability and attack repository for iot," https://data.europa.eu/data/datasets?keywords=variot&locale=en.

[15] "Data extraction laboratory – variot," https://www.variot.eu/2021/10/07/data-extraction-laboratory/.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.