Approaches for IoT Cyber security Analysis Using Various Machine Learning

R. Karthigaichelvi[1], Dr. B. Balakumar[2]

[1]Reseach Scholar, Centre of Information Technology and Engineering, Manonmanium Sundaranar University.

[2]Assistant Professor, Centre for Information Technology and Engineering, Manonmanium Sundaranar University

Abstract: As a result of recent scientific breakthroughs, new technologies are often developed and presented. Since managing and overseeing systems like these can be challenging for humans, our society has turned to machine learning for assistance. New ideas and methods are brought about by new technologies, and new techniques are used to circumvent existing cybersecurity measures. Three alternative Internet of Things (IoT) cyber security algorithms are currently in use in business for malware and intrusion detection: K-Nearest Neighbour (KNN), Support-Vector Machine (SVM), and Random Forest (RF)and. Training and testing were conducted on each algorithm. For malware detection, the highest accuracy of the KNN, SVM, and RF was 90.81%, 84.51%, and 93.37%. For intrusion detection, the highest accuracy was 92.58%, 87.33%, and 93.97%.

Keywords: Internet of Things, KNN, SVM, RF Machine Learning Algorithms, Cyber-Security, Malware, and Intrusion Detections.

1. Introduction

Since 2020, there have been more than 50 billion IoT connections [1]. This provides a significant need to advance the field of cyber security constantly, as stated in [2]. It is necessary to regularly test data security methods with new datasets after they have been reviewed and approved. Innovative and emerging technologies may deviate from those already in use, creating new opportunities for evildoers to take advantage of. To combat DNS-based ad blocking, Google's Home Mini, for instance, uses a hard-coded DNS; however, the result creates a new security concern.Instead of using the local network's DNS settings in this instance, the device uses Google's DNS. With adequate information, a hostile individual can now use the zero-day vulnerability 'Name:Wreck' to exploit the device. Cyber-security experts must be aware of countless situations like this one before an attacker may use them. The efficacy of three various machine learning-based cyber security solutions will be contrasted in this article.We also conduct a study of out-dated machine learning-based IoT cyber security techniques using recently released datasets that reveal previously unseen IoT devices and the technologies/protocols that go along with them, as seen in [3]. We test the SVM, KNN, and RF algorithms that will be utilized in 2021 for IoT cyber security. In Section 2, we will see previous IoT and machine learning algorithms.In Section 3, the dataset and techniques we used in this research are discussed in detail. Section 4 provides and analyzes the outcomes we observed. Section 5 concludes and summarises our research.

2. Earlier IoT applications and Machine Learning Algorithms.

Over two decades have passed since the term "Internet of Things" first appeared. The phrase was first used in 1999 by [4] to describe how radio-frequency identification tags were applied to a production line. The "IoT" is now thought to include any device that can collect data and transmit it over the internet [5]. Because of their versatility and utility, IoT devices are now completely integrated into society's operations.IoT devices are widely used in a variety of settings, including e-commerce, healthcare, and personal living spaces. As mentioned in [6], the Italian city of Padua, for instance, uses IoT networks to monitor a variety of factors, including carbon monoxide levels, traffic flow, noise levels, streetlights, and electricity consumption.Stronger cyber security measures are essential as a result of the sophistication of hostile assaults on these systems increasing along with technological prowess [7]. While there are always new ways to safeguard IoT devices from unwanted users, creating effective cyber security measures specifically for IoT devices presents its own set of particular difficulties.In edge computing contexts, IoT devices are widespread, have similar device signatures, have limited resource availability, and are susceptible to attacks from botnets. Therefore, specialized cyber security is needed that can get beyond these limitations. Studies in [8] focused on the Constrained Application Protocol (CoAP), which provided a useful method for securing real-time data flow within IoT networks.The optimization of this protocol demonstrated that it was possible to achieve in [8] with accuracy rates peaking at 97% which would have otherwise been impossible with current cyber security measures. When compared to other sectors like online transaction fraud prevention, IoT cyber-security might be much enhanced [9].According to [7], the frequency of botnet attacks increases proportionally to the volumetric bandwidth produced by each attack as more internet-connected IoT devices connect. When a Distributed Denial of Service (DDOS) attack with a peak bandwidth of 1.1Tb/s struck the cloud service provider Cloudflare, it was beyond Cloudflare's control and caused significant network outages, demonstrating the power of botnets.SVMs, RFs, and KNNs are the three basic machine learning algorithms utilized by contemporary IoT security solutions for identifying devices, detecting intrusion attempts, and identifying malware. On IoT networks, these three algorithms can essentially serve as watchdogs to prevent assaults in real-time. In this work, we evaluate the applicability of RFs, SVMs, and KNNs for IoT cybersecurity using recent datasets.

To detect the network traffic of IoT devices, [2] suggests the use of supervised learning techniques such as SVM, KNN, RF, Nave Bayes, and Artificial Neural Networks (ANN). In more detail, they can recognize network intrusions and spoofing assaults. Multivariate correlation is required to detect Denial of Service (DoS) attacks.When network traffic variables' geometrical correlations are extracted, the model is 92% more accurate. When deep learning was taken out of the equation, the researchers found that the KNN and RF had the best performance for detecting network intrusion and malware, respectively. Researchers looked in [10] for a machine learning system that might identify DDOS attacks, which were also noted in [11]. They decided to study KNN, SVM, DT, RF, and RF with Linear Kernel. Three IoT devices, 10 minutes of recorded network activity, and the dataset were used to create the dataset.The models were trained using an 85/15 split training technique using the Sci-kit learn Python package. The researchers found that RF performed best and SVM performed worse, and that stateless features were more beneficial for classification than stateful features [10]. Similar to [12], the researchers found that the SVM performed poorly when compared to KNN and RF [13]. They were able to achieve binary classification accuracy levels of 99% using an RF.Additionally, the researchers discovered that the Radial Basis Function kernel for SVM performed nearly twice as well as the Linear kernel. Researchers used Google's MapReduce as a foundation for network traffic feature extraction, translation, and analysis of shifting network features, continuing the pattern from [14]. They investigated seven different machine learning-based algorithms, and RF fared the best with a precision of 0.994, considerably outperforming SVM at 0.775.Five distinct intrusion detection algorithms were put to the test [14] by researchers. They obtained a dataset from Kaggle and chose 8 various feature vectors to use for their machine learning. The training was finished with five-fold cross-validation and 80/20 split testing, as mentioned in [15]. They discovered that RF outperforms SVM, which showed performance regression, with larger datasets. The researchers showed that RFs, Decision Trees, and KNNs could all accurately classify and distinguish between normal and attack data, with an average accuracy rate of around 90.5% for most machine learning algorithms observed in the IoT space.

3. Techniques

Verifying a dataset was published within the year 2021 was the initial stage in the search process. Our analysis determined that the 'Aposemat Iot-23' dataset adequately met all of our requirements after analyzingseveral datasets.Newer gadgets that current cyber security systems haven't yet interacted with are included in freshly published datasets, which can have an impact on the quality of the results. This creates untested conditions and reveals fresh exploits.The dataset was fed into a Pandas framework (v1.2.4) running the Scikit-learn toolkit (v0.24) inside the Spyder IDE (v4.2.3). Holdout validation was used in the study to confirm the accuracy of the trained models for malware identification and intrusion detection. The splits used for training the data were 60/40, 70/30, and 80/20. The averaging of the findings from each split is shown in the ROC curves and confusion matrices that are presented.

To get the maximum performance out of each algorithm, this post optimized the feature sets for both malware and intrusion detection. To begin the process of data optimization, all the features in our labeled traffic capture, shown in Figures 1 and 2, were evaluated.missed_bytes and resp_pkts were two features that were quickly discarded during testing because they were ineffective at detecting malware or intrusion-based attacks. This is because using more basic machine learning methods, these features do not increase the likelihood of detecting malware or intrusion-based attacks.Keep in mind, though, that the inclusion of these attributes may prove advantageous with deep learning algorithms or more sophisticated datasets that can make use of additional data points. Every method used the same basic set of capture features, however some, like RF, performed better when a second feature was added. When the characteristics proto and orig_bytes were included, the RF's malware detection accuracy improved and its false positive rates decreased (Fig. 2).Since malware and non-malicious programs frequently use the same UDP and TCP protocols, the proto feature greatly boosted false positive rates in the KNN and SVM. Additional testing ought to simulate peer-to-peer traffic and MAC addresses that have been anonymized, which is now standard procedure in the cybersecurity industry.

3.1 Detection of malware

Our testing made use of a foundation feature set, as shown in Figure 1, consisting of uid, resp_ip_bytes, resp_pkts, orig_ip_pkts, history, resp_bytes, and duration, to correctly identify malware on the local IoT network. With the use of the uid function, the researchers were able to follow a device across the local network even if its IP changed or identity-masking or spoofing methods were used.When malware is present on the network, the resp_ip_bytes and resp_bytes sizes are often very little or zero. This statistic is useful for malware identification because, in contrast to resp_ip_bytes, certain malware raises the sum of orig_ip_pkts. The history element has very little to no impact on reducing false positive rates. This article preserved it as a feature because it may be used in subsequent testing. After all, the results were not adversely affected by it.The length is also a fantastic tool to use because malware typically has very slow connection times.

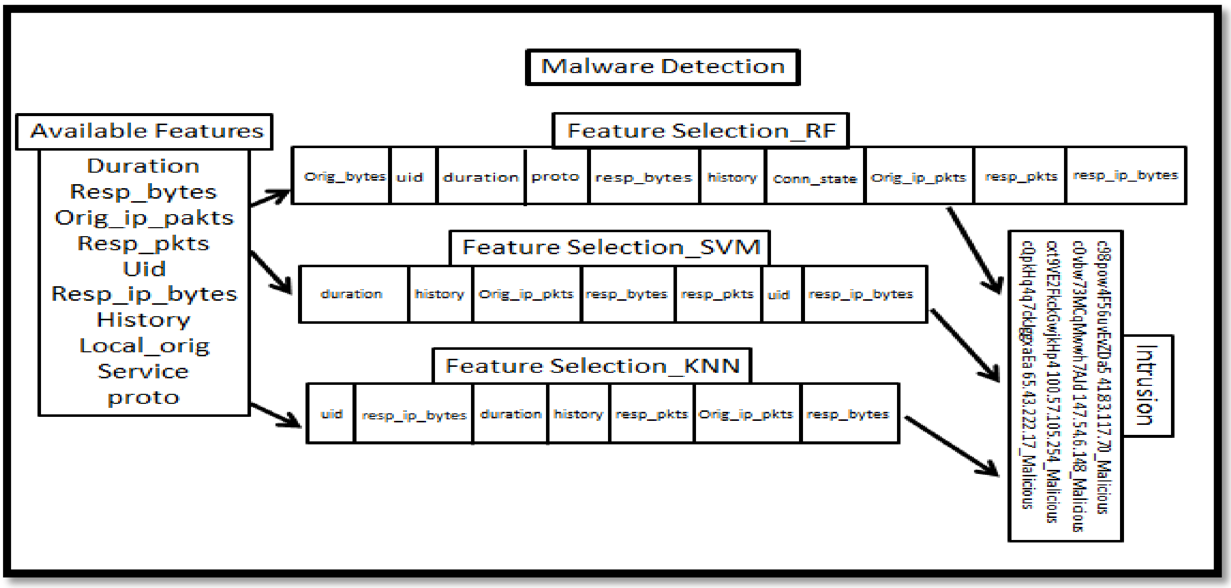

Figure 1 Selection of Features in Malware Detection

3.2 Detection of intrusions

Intrusion detection uses the uid to monitor the device more over the local IoT network, just like virus detection does. They both have the duration feature, which can be used to spot periods that seem suspicious or indicate hostile intent. The timestamps (TS) feature is exclusive to intrusion-based attacks since, depending on the attack type, a vertical or horizontal port scan will have extremely similar TS from the same device over numerous local IPs or ports.It must be associated with an identifier for cross-referencing to make better use of the TS functionality. The IP addresses of the origin device and the destination device can both be found using the id.resp_p and if.orig_h features in this situation. Due to the ability to compare similar timestamps with the host and receiving IP addresses and create meaningful results, this resulted in improved accuracy.

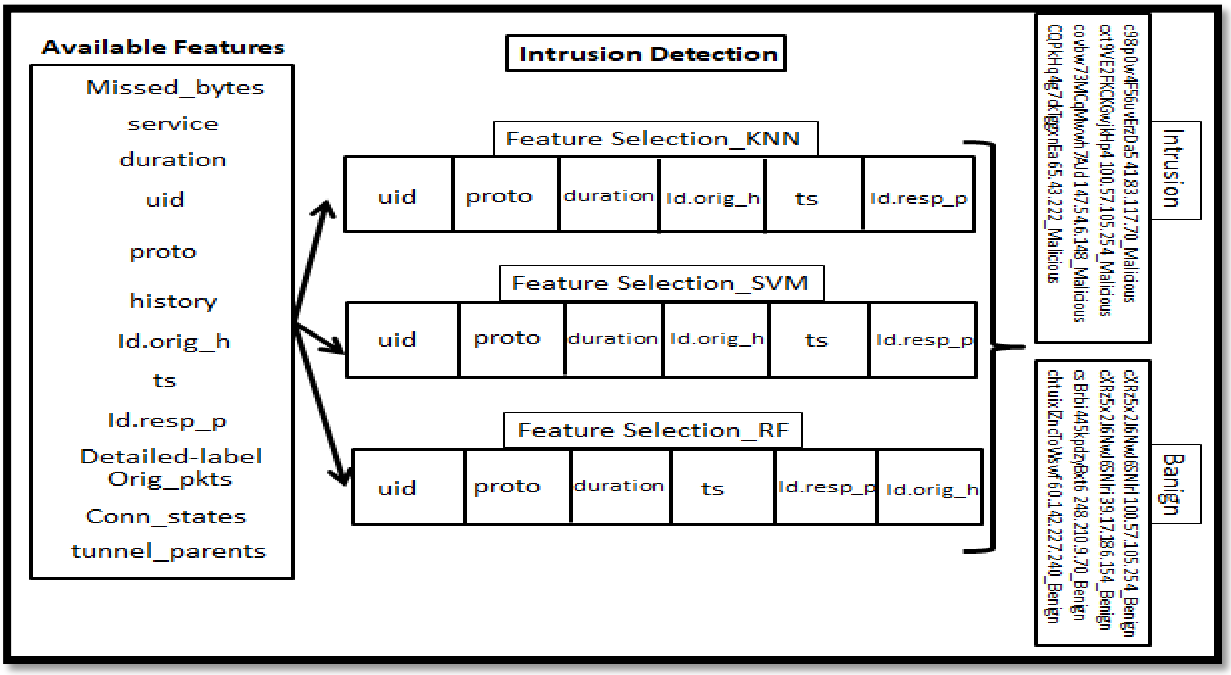

Figure2. Selection of Features in Intrusion Detection

Table. 1 Average for intrusion detection

Algorithm	Accuracy(%)	F1(%)	Recall(%)	TP(%)	FN(%)
RF	93.97	96.89	95.75	96.08	4.19
SVM	87.33	87.81	87.24	87.33	8.37
KNN	92.58	95.28	93.63	92.81	4.67

Table. 2 Averages for Malware Detection

Algorithms	Accuracy	F1	Recall	TP	FN
RF	93.37	94.57	94.30	92.83	4.62
SVM	84.51	86.62	82.41	84.42	6.04
KNN	90.81	93.58	90.81	90.71	5.21

4. Analysis of Results

According to Table 1 of this investigation, the RF algorithm had the best performance. It advertised low overall false positive rates and high accuracy rates for intrusion and malware detection, at 93.97% and 93.37%, respectively. With the Linear kernel for malware detection and the RBF kernel for intrusion detection, it was discovered that the SVM performed more effectively.With F1scores that were within 4% of the RF and accuracies of 92.58% and 90.81%, the KNN was the second-best performing algorithm in terms of intrusion detection and malware detection.We were able to recognize malware on an IoT network as well as horizontal and vertical port scans using the special features chosen in this article.Instead of relying on timely lookups that are subject to human mistakes, this study was able to validate our results for accuracy and false negatives. The authors of the dataset classified the traffic as benign or malicious. The RF, as opposed to the KNN and SVM, showed a more constant ROC curve throughout both scenarios, according to our research.As the SVM revealed unusual inconsistencies, possibly due to the various kernels utilized between malware and intrusion detection, as emphasized in [15], the fundamental causes that may be responsible for this are not immediately understood. This study concluded that despite their age, machine learning algorithms are still useful for detecting malware and intrusions on IoT networks and devices. This was confirmed using a variety of testing and validation techniques. The averaging of the output from these techniques is shown in Tables 1 through 2.The RF method outperformed the other two evaluated algorithms, as found in [2, 16]. Figures 3 and 4 show the ROC curves and confusion matrices for our findings. Our results' visualization enables better comprehension of the data and its level of detail. Higher ROC values show that the algorithm can efficiently discriminate between positive instances in the data, or the context of our experiment, whether a device is benign or malicious. The ROC curve displays the ratio between the true and false positive rates.In our experiment, we saw findings that were comparable to those in [10], in which the authors showed how RFs beat other algorithms and effectively developed a system that can detect malicious or anomalous data in real-time. Additionally, [14] demonstrates this. Similar to our findings, [10] also showed that the RFs function at their best when solely fed stateless feature sets.We successfully recreated [14]'s ability to recognize attack types based on the data and distinguish between horizontal and vertical port scans. This can be seen in our confusion matrices, where all of our algorithms have high true positive rates and low false negative ratings. A more accurate and refined conclusion can be reached by concluding that each technique is appropriate for detecting recent malware and intrusions.

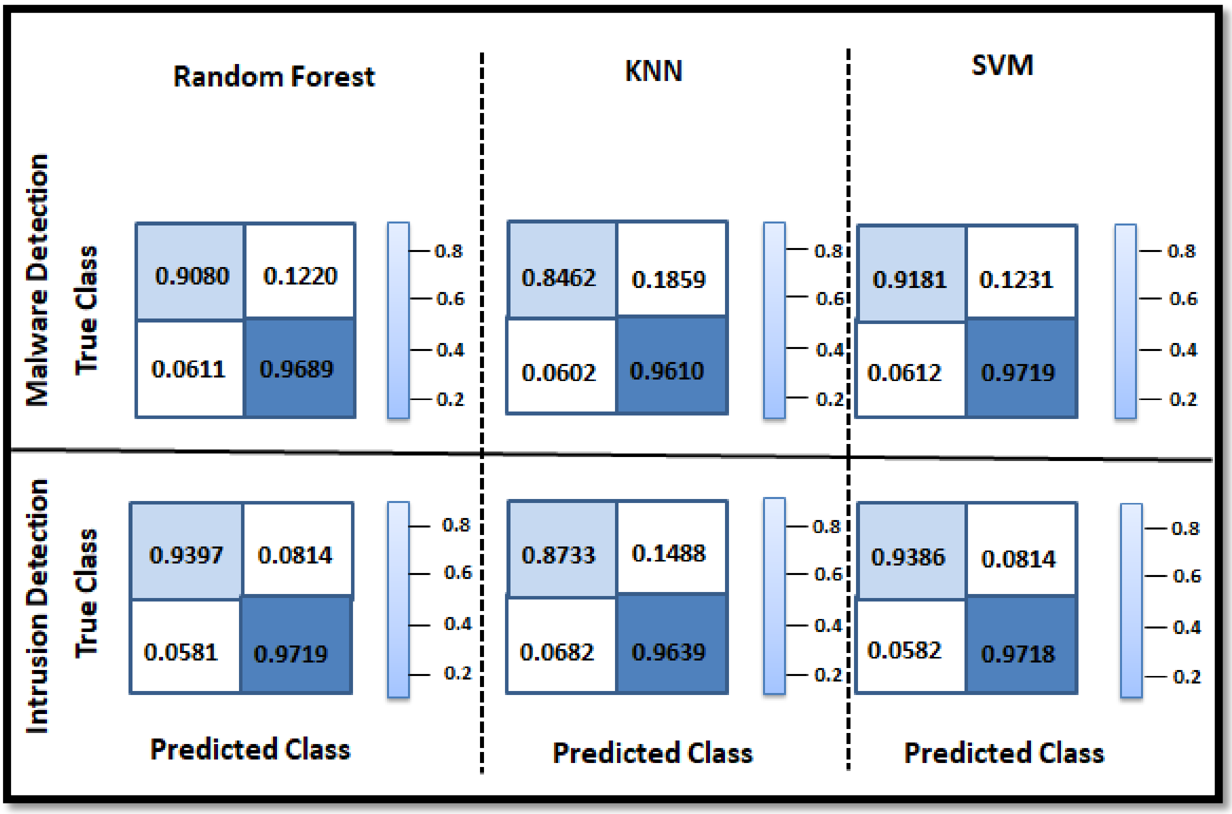

Figure. 3 Comparison of Confusion Matrix with Different Algorithms

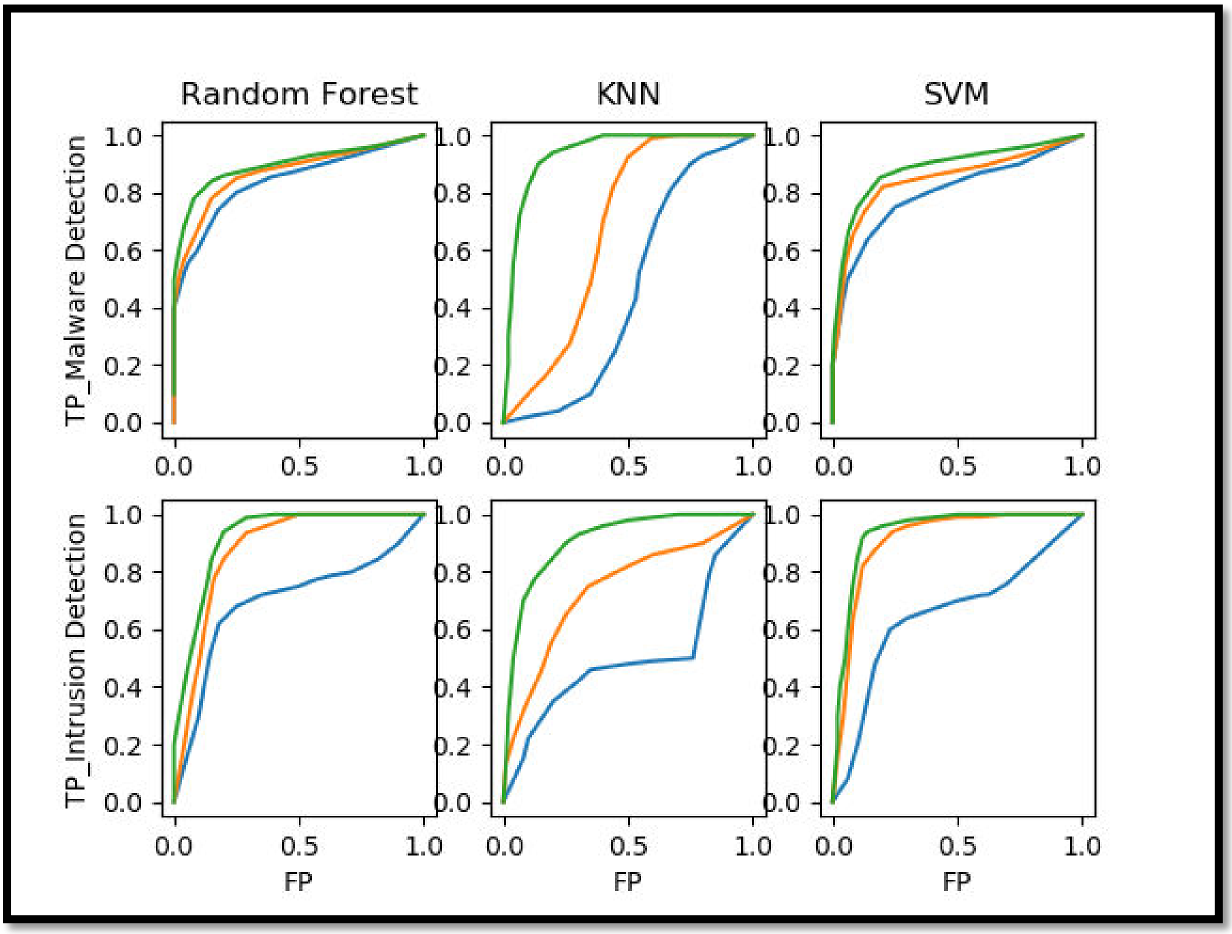

Figure 4 Comparison of ROC Curve with diverse Algorithms

5. Conclusion

This article made clear how crucial it is to continue to develop and maintain cyber security procedures for IoT networks and devices to thwart harmful attackers. Our study's limitations include the need to test more SVM kernels, such as polynomial kernels, to improve SVM outcomes.Future kernel testing, in our opinion, may result in far higher SVM for malware detection accuracy. This might help fix the ROC curve issues highlighted before.Combining datasets to enable more variation in networking setups, devices, and protocols to allow for additional factors to be tested against would be a minor but perhaps missed enhancement for future studies. Additionally, bigger datasets might depict the entropy in real-world systems more precisely.The dataset employed in this article's analysis's age-based limitations is another factor. Despite the dataset's 2020 publication, some of the captures contained within go back as far as 2018. Due to this, the problem of trained models lagging up to three years behind present IoT networks is raised.A three-year time frame provides for a variety of IoT device changes, including firmware upgrades, OS updates, hardware revisions, and new protocols. All of these changes might potentially lead to attacks, necessitating on-going monitoring of cybersecurity measures to preserve their efficacy.This paper examined the effectiveness of RFs with contemporary IoT data and found that it is still the best algorithm to employ for IoT cybersecurity. While the KNN and SVM did not outperform the RF in terms of performance, they can still be utilized successfully in today's IoT networks to detect malware and intrusions.

Reference

1. Jayakumar, H., Lee, K., Lee, W. S., Raha, A., Kim, Y., &Raghunathan, V. (2014). Powering the internet of things. In Proceedings of the 2014 international symposium on Low power electronics and design (pp. 375-380).
2. L. Xiao, X. Wan, X. Lu, Y. Zhang and D. Wu, (2018) "IoT Security Techniques Based on Machine Learning: How Do IoT Devices Use AI to Enhance Security?," in IEEE Signal Processing Magazine, vol. 35, no. 5, pp. 41-49,
3. Hindy, H., Bayne, E., Bures, M., Atkinson, R., Tachtatzis, C., &Bellekens, X. (2020, September). Machine learning based IoT Intrusion Detection System: an MQTT case study (MQTT-IoT-IDS2020 Dataset). In International Networking Conference (pp. 73-84).Springer, Cham.
4. Davis, G. (2018). 2020: Life with 50 billion connected devices. 2018 IEEE International Conference on Consumer Electronics (ICCE).
5. Gunn, Dylan J. et al. (2019) "Touch-Based Active Cloud Authentication Using Traditional Machine Learning and LSTM on a Distributed Tensorflow Framework." *Int. J. Comput. Intell. Appl.* 18 1950022:1-1950022:16.
6. Miorandi D, Sicari S, De Pellegrini F, Chlamtac I (2012) Internet of things: vision, applications and research challenges. Ad Hoc Netw 10(7):1497.
7. T. Kelley and E. Furey, (2018), "Getting Prepared for the Next Botnet Attack: Detecting Algorithmically Generated Domains in Botnet Command and Control," 2018 29th Irish Signals and Systems Conference (ISSC), Belfast, pp. 1-6,
8. Shelton, J., Rice, C., Singh, J., Jenkins, J., Dave, R., Roy, K., &Chakraborty, S. (2018, August). Palm Print Authentication on a Cloud Platform. In 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD) (pp. 1-6).IEEE.
9. Mason, J., Dave, R., Chatterjee, P., Graham-Allen, I., Esterline, A., & Roy, K. (2020, December). An Investigation of Biometric Authentication in the Healthcare Environment.Array, 8, 100042.
10. Doshi, R., Apthorpe, N., &Feamster, N. (2018). Machine Learning DDoS Detection for Consumer Internet of Things Devices.IEEE Security and Privacy Workshops (SPW).
11. Cenedese, A., Zanella, A., Vangelista, L., &Zorzi, M. (2014). Padova Smart City: An urban Internet of Things experimentation. Proceeding of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks 2014.
12. Alrashdi, Ibrahim, et al. (2019)"Ad-iot: Anomaly detection of iotcyberattacks in smart city using machine learning." IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). IEEE
13. Chhabra, Gurpal Singh, Varinder Pal Singh, and Maninder Singh. (2020): "Cyber forensics framework for big data analytics in IoT environment using machine learning." Multimedia Tools and Applications 79.23 15881-15900.
14. Hasan, M., Milon Islam, M., Islam, I., &Hashem, M. M. A. (2019). Attack and Anomaly Detection in IoT Sensors in IoT Sites Using Machine Learning Approaches. Internet of Things, 100059
15. Moh, M., &Raju, R. (2018) Machine Learning Techniques for Security of Internet of Things (IoT) and Fog Computing Systems. 2018 International Conference on High Performance Computing & Simulation (HPCS).
16. Meidan, Y., Bohadana, M., Shabtai, A., Ochoa, M., Tippenhauer, N. O., Guarnizo, J. D., &Elovici, Y. (2017). Detection of unauthorized IoT devices using machine learning techniques.arXiv preprint arXiv:1709.04647.

.