Exploring Machine Learning Models

For IoT Network Intrusion Detection: A Literature Review and Comparative Analysis

Jack Hagen, Ayden Strate | Advisor: Dr. Mounika Vanamala | Computer Science

University of Wisconsin Eau Claire

Introduction

The Internet of Things (IoT) is a complex network of low-powered interconnected devices that communicate and exchange data over the internet. With interconnectivity, there comes the significant challenge of vulnerability to cyber threats. IoT devices are a popular target for cybercriminals because of the sensitive user data they handle and the lack of security updates most IoT devices receive. As cyber defenses mature, so does the complexity of cyber-attacks. Sophisticated attackers can stay undetected inside of a network for long amounts of time because of the complexity of modern networks. Traditional intrusion detection methods, usually reliant on monitoring by network specialists, struggle to keep pace with the evolving threat landscape. Although this is true, and Machine Learning based Intrusion Detection Systems (IDS) show promise, they are not as effective as traditional manual monitoring by network security specialists. Through this, Machine Learning holds the potential to enhance detection capabilities by autonomously identifying anomalous activities, even with current implementations being behind in effectiveness of human-led monitoring. In response to this difference in effectiveness, this research leveraged several existing network activity databases and their trained Machine Learning models to identify novel machine learning algorithms tailored to improve the efficacy of intrusion detection. It aims to implement new machine learning algorithms with existing network activity databases to improve intrusion detection effectiveness. The beginning of this implementation starts with this literature review on simulations of popular IoT cyber-attacks within virtual environments. These databases include attributes of a live network, including both malicious and benign connections. They contain rich metadata which allows ML models to classify connections as potentially malicious. Overall, this research conducted a literature review of ten studies including twenty-eight algorithms and nine datasets. These studies were compared by datasets to algorithms, to see which algorithms should be the primary considerations for further, in-depth study.

Research

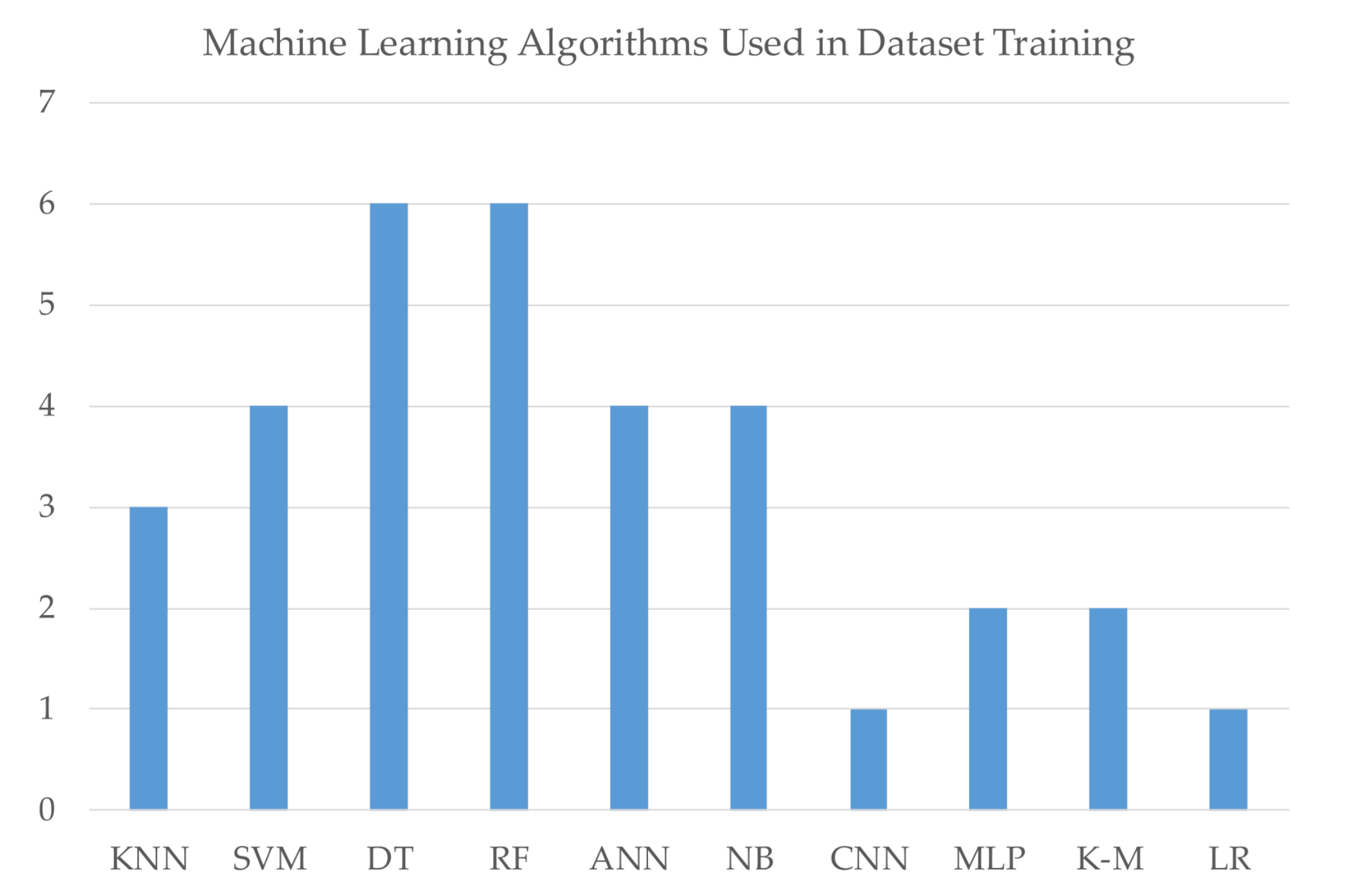
Datasets

Each dataset corresponds to an individual study conducted; these studies supported a large variety of training and analysis methods, of which we went through and organized a table that considers which datasets have been investigated by which algorithms. Below is a list of datasets considered under this literature review.

- CICIDS 2017
  - The authors built a profile-based system with benign background data with a mix of modern attacks (Brute Force FTP, Brute Force SSH, DoS, etc.).
- UNSW-NB15
  - Gathered real network data from the cyber lab and contained nine different attack types (Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms).
- DS2OS
  - Contains regular activity of a network of IoT devices. Gathered over one day at the application layer.
- BoT-IoT
  - The authors created realistic networks at the Cyber Range and exposed it to DDoS, DoS, OS and Service Scan, Keylogging and Data exfiltration attacks. The network contained normal and botnet traffic.
- TON-IoT
  - Data collected from a synthetic network designed that incorporates attack data from DoS, DDoS and ransomware, against web applications, IoT gateways and computer systems.
- KDD CUP 1999
  - Entry for The Third International Knowledge Discovery and Data Mining Tools Competition. Contains network traffic from a simulated military network.
- NSL-KDD
  - Created by the same organization, tried to solve the shortcomings of KDD1999.
- ISCX2012
  - The authors then programmed tools to emulate users based on abstractions, mimicking real user activity. The authors then designed and implemented attack strategies.
- ANDROzoo
  - A collection of publicly available APK files. All or approaching all are scanned for malware.

Dataset Algorithm Matrix

Dataset	Algorithms	Key
CICIDS 2017	KNN, SVM, DT, RF, ANN, NB, CNN, K-M, EM	KNN – K Nearest Neighbor SVM – Support Vector Machine DT – Decision Tree RF – Random Forest ANN – Artificial Neural Network NB – Naive Bayes CNN – Convolutional Neural Network MLP – Multilayer Perception K-M – K Means Clustering E-M – Expectation-Maximization Clustering LR – Logistic Regression
UNSW-NB15	SVM, NB, DT, RF	
BoT-IoT	SVM, LR, DT, RF, ANN, KNN, MLP ANN, NB	
TON-IoT	XGBoost, LR, RF, DT	
KDD CUP 1999	MLP, GAU, K-M,DT	
NSL-KDD	KNN DT, RF, SVM, ANN	
ISCX2012	SVM, RBF, RF, KNN	
ADROzoo	RF, SVM, LR, DT, XGBoost, KNN	


Machine Learning Algorithms Used in Dataset Training

Literature Review

Two Studies

Through this collection of datasets, there were two separate studies conducted during this literature review process. The first study included a consideration of another IDS systematic literature review which provided a list of 49 investigations that used deep learning and various machine learning algorithms. The second study was conducted with a similar approach, but included describing the datasets themselves, and within, what sources of data were considered for collection. With this additional consideration came the further ability to determine the depth of investigation.

Results

Through these two studies in the literature review process, we noticed that they used multiple algorithms to improve detection rates. We also noticed that the datasets use relatively outdated data, the latest recorded dataset was created in 2020. As network architecture and software changes, it's important that datasets in use are based off the latest data to more accurately classify network behavior. Overall, this literature review provided further context to not only the fast speed at which this technology is developing, but it illustrated the importance of testing several algorithms upon a single dataset, as many of the studied investigations found improvement when more algorithms were involved.

References

Rafiq, H., Aslam, N., Aleem, M. et al. AndroMalPack: enhancing the ML-based malware classification by detection and removal of repacked apps for Android systems. Sci Rep 12, 19534 (2022). https://doi.org/10.1038/s41598-022-23766-w

Injadat, MohammadNoor, et al. "Bayesian Optimization with Machine Learning Algorithms Towards Anomaly Detection." arXiv, arXiv:2008.02327v1 [cs.LG], 5 Aug 2020.

Mari, Andrei-Grigore et al. “Development of a Machine-Learning Intrusion Detection System and Testing of Its Performance Using a Generative Adversarial Network.”Sensors (Basel, Switzerland) vol. 23,3 1315. 24 Jan. 2023, doi:10.3390/s23031315

Sabhnani, Maheshkumar. "Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misus Detection Context." EECS Dept, University of Toledo, Toledo, Ohio 43606 USA

Karanfilovska, Martina, et al. "Analysis and modelling of a ML-based NIDS for IoT networks." Procedia Computer Science, vol. 214, 2022, pp. 570-577. https://doi.org/10.1016/j.procs.2022.08.023.

Z. K. Maseer, R. Yusof, N. Bahaman, S. A. Mostafa and C. F. M. Foozy, "Benchmarking of Machine Learning for Anomaly Based Intrusion Detection Systems in the CICIDS2017 Dataset," in *IEEE Access*, vol. 9, pp. 22351-22370, 2021, doi: 10.1109/ACCESS.2021.3056614

Le, D. Q., Nguyen, P. Q., & Truong, N. B. “Intrusion Detection in IoT Networks: A Machine Learning Perspective” Proceedings of the International Conference on Computing, Communication and Wireless Systems (ICCWS 2022)

M. -O. Pahl and F. -X. Aubet, "All Eyes on You: Distributed Multi-Dimensional IoT Microservice Anomaly Detection," 2018 14thInternational Conference on Network and Service Management (CNSM), Rome, Italy, 2018, pp. 72-80

Pokhrel, S., Abbas, R., Aryal, B. "IoT Security: Botnet Detection in IoT using Machine Learning.“ arXiv, arXiv:2104.02231 2021

Rafiq, H., Aslam, N., Aleem, M. et al. AndroMalPack: enhancing the ML-based malware classification by detection and removal of repacked apps for Android systems. Sci Rep 12, 19534 (2022). https://doi.org/10.1038/s41598-022-23766-w

Conclusion

The surveyed datasets include data which can be used to detect malicious network activity, but many of the datasets are outdated and would likely not be able to fit in the modern network landscape.

Future Work

Based on the insights gained from our research, there is an opportunity to explore the application of several algorithms that have not yet been utilized on the TON-IoT dataset. This study considered in the literature review contained a very recent NetFlow dataset iteration under the name NF-TON-IoT-v2, which has potential for a helpful understanding of network traffic patterns and anomalies in IoT environments

We thank the Office of Research and Sponsored Programs for supporting this research, and Learning & Technology Services for printing this poster.