Military Cyber Affairs

Volume 7
Issue 1 *HammerCon 2024 Issue*

Article 4

May 2024

Generative Machine Learning for Cyber Security

James Halvorsen
Washington State University, james.halvorsen@wsu.edu

Dr. Assefaw Gebremedhin
Washington State University, assefaw.gebremedhin@wsu.edu

Generative Machine Learning for Cyber Security

James Halvorsen and Dr. Assefaw Gebremedhin

Introduction

The past decade has seen significant improvements in the development of generative models. Tools such as Stable Diffusion (Rombach et al., 2022) and ChatGPT (OpenAI, 2023) have become household names within a short period of time given their ability to transform how people get creative work done. A lesser-known capability of generative models, though still of significant importance, is their ability to improve our nation's cyber security infrastructure.

To someone unfamiliar with both machine learning and cyber security, the relevance of generative models to cyber security applications may not be obvious. We summarize this relevance in a few short points:

1. To defend against unknown and creative adversaries, we must have effective automated tools that can detect and respond to attacks on our networks.
2. Current applications for this task suffer from high false positive rates.
3. Developing and testing more effective defenses is difficult due to lack of quality data, which can be expensive to produce.
4. Generative models are effective at improving low quality datasets.
5. Some generative models are additionally effective at developing new intrusion detection software.

This paper expands upon each of the above points in greater detail, with the overall goal of demonstrating the importance of generative models for the future of cyber security. It is organized as follows. Section II provides an overview of existing problems related to cyber security and how machine learning has been used thus far to combat them. Section III demonstrates the relative strengths of generative models in improving the current situation. Section IV covers some of the current weaknesses with generative models with respect to cyber security problems and discusses where further research on this subject is needed. Section V concludes the paper.

Overview of Machine Learning and Cyber Security Issues

According to a Statistia (2023) report, the cost of cyber attacks to businesses and governments alike has increased dramatically over the past several years and is estimated to increase even more in the coming years. The causes for this are numerous. To start, the simultaneous development of anonymous cryptocurrencies and ransomware have enabled criminal organizations to extract fees from corporate networks while remaining undetected, creating an incentive to engage in criminal activity (August et al., 2022). Additionally, there is an increasing level of connectivity between critical infrastructures, such as electrical grids and water distribution networks, and cyber infrastructures (De Brujin et al., 2017). Given the success of previous attacks against critical infrastructures, such as Stuxnet (Baezner et al., 2017) and BlackEnergy (Geiger et al., 2020), we can expect further investment by state actors in their capacity to carry out future attacks of this kind.

With this increase in desire and capacity to carry out cyber attacks, cybersecurity professionals must in turn increase their capacity to defend against them. Given the sheer quantity of organizations that have networks to defend, this need cannot be met by human labor alone. Automation thus offers an alternative solution to meeting our defensive requirements. Further, since the cyber threat landscape is both complex and regularly changing, these automated solutions will need to use machine learning to successfully adapt to the novel attacks they will be encountering on a regular basis.

Intrusion Detection Systems (IDS) are a class of applications best suited for automating the security of our nation's cyber infrastructure. First proposed by Anderson (1980), IDSs use statistical models of some subject under observation (such as a host or network), to make inferences about whether that subject is experiencing a cyber attack. A human administrator can use these inferences to form a response to the attack. Since their invention, the capabilities and design of IDSs have been significantly expanded using machine learning techniques (Ahmad et al., 2021). Considering that this amounts to several decades of research into refining our abilities to detect and potentially even prevent cyber attacks, we are left with an important question: *why are defenders still losing the cyber arms race?*

A key part of the answer to this question lies in a flaw with many IDSs that limits their widespread use. That flaw is false positives (Kizza, 2024; Markevych et al., 2023). Most traffic on any given network is benign, and a portion of all benign traffic will always be anomalous. Learning to distinguish between behavior that is uncommon and benign, and behavior that is malicious, is a particularly difficult problem that is made even more difficult by a separate issue, namely, lack of adequate data.

Data is a key component of any application that uses machine learning, whether it is using supervised learning, unsupervised learning, or any other approach. In the context of intrusion detection, a good dataset should be labeled,

have a variety of different types of attacks, be balanced (i.e., every class should have a reasonable number of samples), and be recent. Many public datasets for intrusion detection lack one or more of these attributes (Małowidzki, 2015).

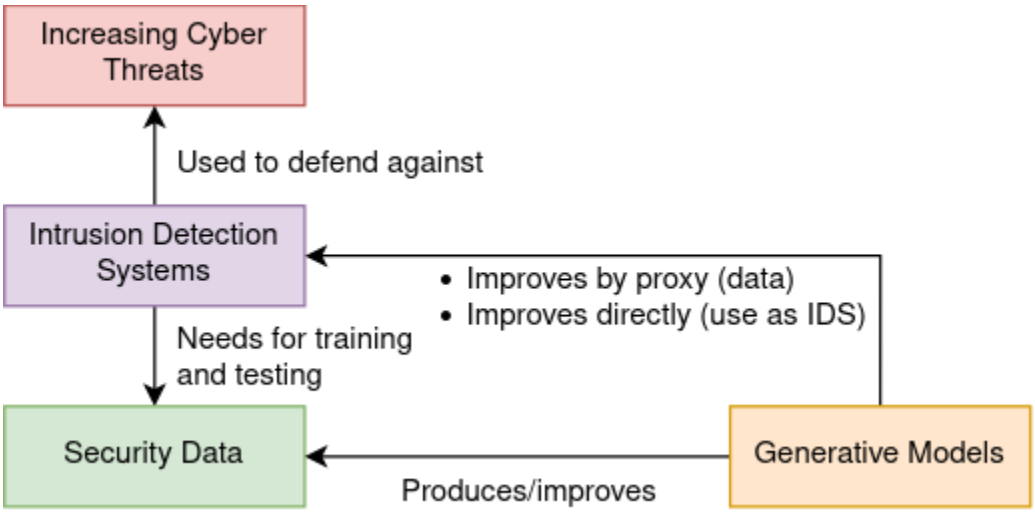

Figure 1. A hierarchy of how data issues relate to cyber threat issues, and how generative models may be used to alleviate problems.

These data-related issues are impediments to intrusion detection research that indirectly make combating future cyber threats difficult. The relationship between these issues is depicted in Figure 1, along with how generative models stand to alleviate the issues. Ultimately, solving the problem of increased cyber threats requires both improving IDS performance and addressing the problems related to data.

Strengths of Generative Models

Generative models offer several solutions to the issues related to machine learning and cyber security discussed in Section II. In our review of the literature, we identify three application areas for generative models to improve cyber security and discuss them in this section. The first application is improving cyber security datasets which are unbalanced by generating new samples of minority classes within the dataset. This allows for improved training of IDSs. A second application of generative models is to generate synthetic adversarial samples designed to uncover weaknesses within existing defenses. This allows for improved testing of network defenses. Finally, the architecture of certain generative models (namely Generative Adversarial Networks, or GANs, and Variational Autoencoders, or VAE) allows for training of new types of IDSs that can perform better than traditional deep learning approaches.

Improving datasets with generative models is perhaps their most intuitive use. While generative models cannot create cyber security data from scratch, they can transform a dataset that is ineffective for IDS training into one that is much more effective. This is accomplished by increasing the relative proportion of attack samples to benign samples within the dataset so that IDSs trained on that data do not become victims of overfitting.

Empirical support for the effectiveness of generative models for this application is provided by Merino et al. (2020) and Yilmaz et al. (2020). Merino et al. (2020) used a GAN to produce synthetic samples of minority classes within the KDD99 dataset (Stolfo et al., 1999), and when evaluated using a classifier, 100% of the synthetic samples were correctly identified as attacks. Yilmaz et al. (2020) performed a similar task with GANs on the UGR'16 dataset, creating enough attack samples to match the number of benign samples. Across seven attack classes, each having less than 100 samples prior to the introduction of synthetic samples, classifier performance trained on the resulting synthetic datasets improved in terms of precision and recall. In this evaluation, the accuracy was also improved slightly, although this improvement comes from a position where a classifier trained on the original, unmodified dataset could attain an accuracy of 99% by classifying all samples as benign.

The weakness of an unbalanced dataset is that correct identification of minority classes (i.e., attack data) is not necessary to achieve high accuracy. In each of these works, the role of generative models is to make correct identification of these classes a requirement, which ultimately results in better classification models.

Beyond improving IDS training, testing cyber defenses is another important capability of generative models. This capability can manifest in the form of a variety of different simulated attacks. Ahmadian et al. (2018) used a GAN to conduct a false data injection (FDI) attack against a simulated smart grid environment, demonstrating that this could be used to manipulate energy prices for a profit-seeking attacker. GANs can also be used for exploring a target's defenses. Shi et al. (2018) demonstrate this by using a GAN to attack a machine learning classifier with limited API access and a model that is hidden from the attacker. The GAN would generate samples for the classifier to give labels to and use them to train an equivalent classifier. If this classifier were an IDS, duplicating its model in this manner would enable an attacker to learn what attacks would be incorrectly classified, and use this information to craft an attack that the IDS will not be able to detect.

These specific capabilities of generative models have applications that are both offensive and defensive. A defender may be interested in using generative models to learn the weaknesses in their defenses, while an attacker may be interested in using generative models to learn the weaknesses in their victim's defenses. While researching this topic further may seem to be a double-edged

sword, these capabilities are already published in public research. What we do not research, our adversaries will be certain to take advantage of.

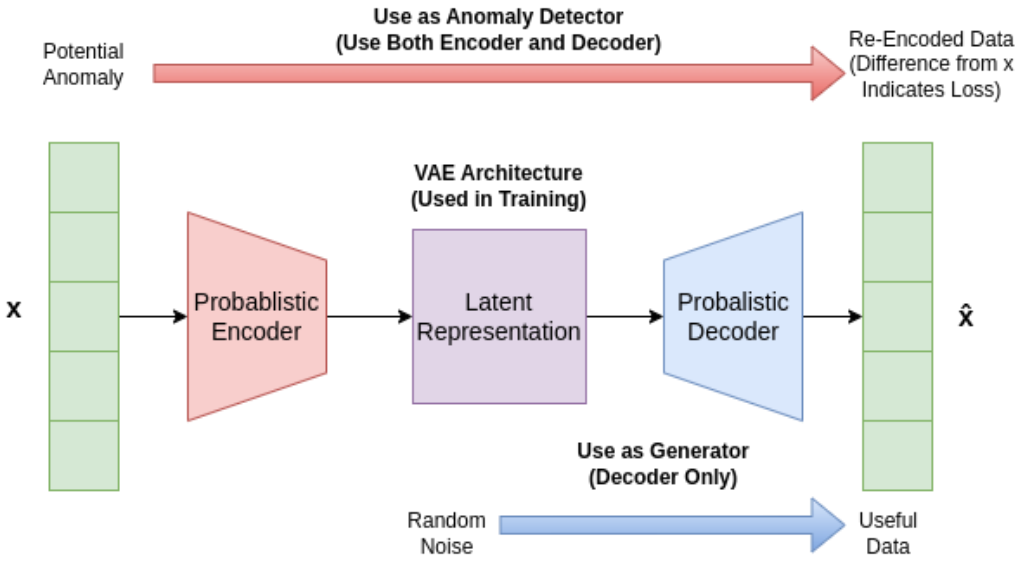

Figure 2. Architecture of VAE as used for generative and anomaly detection tasks.

Thankfully, most applications of generative models in cyber security that we have observed appear to be purely defensive in nature. Two architectures of generative models, GANs and VAEs, are designed in such a way that they can perform tasks other than generating data, which can be enhanced by their generative models. GANs contain two neural networks trained simultaneously, one for generating data and one for determining whether the generated data is synthetic. This second model, known as a discriminator, can be modified to be used for anomaly detection purposes (Jan et al, 2018). VAEs also contain two neural networks trained simultaneously, although neither performs a classification task on their own. Rather, one is used to encode data into a latent representation, and the other decodes the latent representation into its original form, sometimes with a reconstruction error. This reconstruction error from using both the encoder and decoder together can also be used for anomaly detection. Figure 2 demonstrates the difference in how its architecture can be used for both generative tasks and anomaly detection.

Both strategies can perform quite well. Jan et al. (2018) showcase this with a Deep Convolutional GAN (DCGAN) that has been modified for detecting Android malware. This DCGAN model is compared to several other machine learning models and demonstrates a relatively high performance on the same dataset. Most notably, however, its false positive rate of only 0.2% shows significant potential for GAN-based models in future IDS development. Zavrak et al. (2020) show that VAEs have similar superiority over traditional classifier models when detecting several types of network-based attacks.

Areas for Improvement

While generative models provide significant contributions to the training and testing of IDSs, there are still several topics related to generative models that need continued research for their maximum potential to be realized within this domain.

Among the greatest challenges for applying generative models to cyber security is creating standard *feature representations* (Ring et al., 2019). Generative models have shown the greatest level of success in image generation tasks, where the data to be generated already has continuous features. Mapping pixel colors to IEEE floating point numbers used in a feature vector is trivial, as is computing loss. If a generator is meant to produce a red flower, and instead produces a magenta one, the distances between the values in the red, blue, and green color channels will be much lower than if it produced a solid blue flower. By contrast, consider how one might encode a TCP port in a netflow record. If port 22 (used for SSH) would be correct for a record, generating port 21 (FTP) is just as wrong as generating port 6667 (IRC). Though port numbers are just integers, a simple mapping to continuous values will not work.

Ring et al. (2019) have done extensive research on finding adequate feature representations for netflow, which contain several different types of categorical features, such as ports, IP addresses, and transport protocols. Netflows are not the only type of data used in cyber security tasks, however. Consider the task of generating packet captures: every protocol could require a different method of representing data. Generating data for host-based intrusion detection will have similar problems, as features may include file paths or payloads within files. Future research relating generative models to cyber security will need to consider how to create feature representations for a wide variety of security data.

Another challenge of note concerns *metrics* (Betzalel et al., 2022). A common metric used for measuring the quality of generated data is called “Train on Synthetic, Test on Real” (TSTR) (Zingo et al., 2021). This process involves using a classifier trained on data produced by the generative model to predict class labels in a dataset containing non-synthetic samples. The advantage of this method is that it allows synthetic data to be evaluated in its intended purpose of predicting real traces of cyber attacks. However, it is also difficult to use this method to compare different works, even on the same original dataset, as they may use a different classifier.

Metrics that use the statistical distribution of datasets, such as the Fréchet Inception Distance (FID) are also common. Although they can be compared between works, the elements of quality being emphasized by comparing the synthetic and real distributions in this manner differs substantially from a classifier-

based approach. In particular, one must consider that a significant goal of using generative models in cyber security is to change the overall distribution of a dataset so that minority classes (i.e., attack data) are more strongly represented. This may make metrics focused on statistical distributions less effective in measuring overall data quality. Future research should consider alternative metrics that may better measure data quality than either statistical or classifier-based approaches.

Conclusion

The scale of current and future cyber threats demands effective machine learning tools for effective cyber defense. Research into generative machine learning has shown that generative models stand to provide numerous benefits to current machine learning tools used for cyber security applications, such as IDSs. These benefits cover a broad range of factors involved in the development of machine learning tools, including both their training and testing.

There are, however, some challenges unique to cyber security where generative machine learning is concerned. These challenges are primarily concerned with feature representation of cyber security data, and evaluation of synthetic data quality. Continued research into generative machine learning should address these challenges, thereby improving the contributions of generative models to tackling growing cyber security threats.

About the Authors

James Halvorsen is a Ph.D. candidate in Computer Science at Washington State University (WSU) whose research lies at the intersection of cyber security and machine learning. His current work focuses on applying new generative machine learning techniques to the creation of synthetic cyber security data. He is involved with the VICEROY Cybersecurity Education and Research (CySER) Institute as a graduate mentor and is a Fellow of a Graduate Assistance in Areas of National Need (GAANN) program at WSU.

Dr. Assefaw Gebremedhin is the Lead Principal Investigator for the VICEROY CySER Institute, Director of the GAANN program at WSU, and currently an associate professor with the School of Electrical Engineering and Computer Science at WSU, where he leads the Scalable Algorithms for Data Science Laboratory. His research interests include data science and AI, network science, high performance computing, and cyber security.

Acknowledgement

This publication was supported by Award Number SA10012022020481 from the Griffiss Institute for the VICEROY program. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the sponsor.

References

Ahmad, Zeeshan, Adnan Shahid Khan, Cheah Wai Shiang, Johari Abdullah, and Farhan Ahmad (Jan. 2021). "Network Intrusion Detection System: A Systematic Study of Machine Learning and Deep Learning Approaches". In: Transactions on Emerging Telecommunications Technologies 32. Doi: 10.1002/ett.4150.

Ahmadian, Saeed, Heidar Malki, and Zhu Han. (2018). "Cyber Attacks on Smart Energy Grids Using Generative Adversarial Networks." In 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 942–946.

Anderson, James P (1980). "Computer Security Threat Monitoring and Surveillance". In: Technical Report, James P. Anderson Company.

August, Terrence, Duy Dao, and Marius Florin Niculescu (2022). "Economics of Ransomware: Risk Interdependence and Large-Scale Attacks". In: Management Science 68.12, pp. 8979–9002. Doi: 10.1287/mnsc.2022.4300. Preprint: https://doi.org/10.1287/mnsc.2022.4300. url: https://doi.org/10.1287/mnsc.2022.4300.

Baezner, Marie and Patrice Robin. (2017). Stuxnet. ETH Zurich, 2017. doi: 10.3929/ETHZ-B-000200661. Available: http://hdl.handle.net/20.500.11850/200661

Betzalel, Eyal, Coby Penso, Aviv Navon, and Ethan Fetaya. (2022). "A Study on the Evaluation of Generative Models." arXiv, 2022. doi: 10.48550/ARXIV.2206.10935. Available: https://arxiv.org/abs/2206.10935

De Bruijn, Hans, and Marijn Janssen. (2017). "Building Cybersecurity Awareness: The Need for Evidence-Based Framing Strategies." Government Information Quarterly 34, 1 (2017), 1-7. https://doi.org/10.1016/j.giq.2017.02.007 Open Innovation in the Public Sector.

Geiger, Marcus, Jochen Bauer, Michael Masuch, and Jörg Franke. (2020) "An Analysis of Black Energy 3, Crashoverride, and Trisis, Three Malware Approaches Targeting Operational Technology Systems." 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE, Sep. 2020. doi: 10.1109/etfa46521.2020.9212128. Available: http://dx.doi.org/10.1109/ETFA46521.2020.9212128

Jan, Salman, Shahrulniza Musa, Toqeer Syed, and Ali Alzahrani. (2018). "Deep Convolutional Generative Adversarial Networks for Intent-based Dynamic Behavior Capture." International Journal of Engineering and Technology 7 (12 2018), 101–103.

Kizza, Joseph Migga. (2024). System Intrusion Detection and Prevention. Texts in Computer Science. Springer International Publishing, pp. 295–323, 2024. doi: 10.1007/978-3-031-47549-8_13. Available: http://dx.doi.org/10.1007/978-3-031-47549-8_13

Małowidzki, Marek, Przemyslaw Berezinski, and Michał Mazur (Apr. 2015). "Network Intrusion Detection: Half a Kingdom for a Good Dataset". In: Proceedings of NATO STO SAS-139 Workshop, Portugal.

Markevych, Michal and Maurice Dawson. (2023). "A Review of Enhancing Intrusion Detection Systems for Cybersecurity Using Artificial Intelligence (AI)." International conference Knowledge-Based Organization, vol. 29, no. 3. Walter de Gruyter GmbH, pp. 30–37, Jun. 01, 2023. doi: 10.2478/kbo-2023-0072. Available: http://dx.doi.org/10.2478/kbo-2023-0072

Merino, Tim, Matt Stillwell, Mark Steele, Max Coplan, Jon Patton, Alexander Stoyanov, and Lin Deng. (2020). Expansion of Cyber Attack Data from Unbalanced Datasets Using Generative Adversarial Networks. Springer International Publishing, Cham, 131-145. https://doi.org/10.1007/978-3-030-24344-9_8

OpenAI (2023). GPT-4 Technical Report. arXiv: 2303.08774 [cs.CL].

Ring, Markus, Daniel Schlör, Dieter Landes, and Andreas Hotho. (2019). "Flow-Based Network Traffic Generation Using Generative Adversarial Networks." Computers & Security 82 (2019), 156-172. https://doi.org/10.1016/j.cose.2018.12.012

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer (2022). "High-resolution image synthesis with latent diffusion models." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684-10695.

Shi, Yi, Yalin E. Sagduyu, Kemal Davaslioglu, and Jason H. Li. (2018). "Generative Adversarial Networks for Black-Box API Attacks with Limited Training Data." 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) (2018), 453-458

Statista (Sept. 2023). Estimated Cost of Cybercrime Worldwide 2017-2028 (in Trillion U.S. Dollars). url: https://www.statista.com/forecasts/1280009/cost-cybercrime-worldwide.

Stolfo, Salvatore, Wei Fan, Wenke Lee, Andreas Prodromidis, and Philip Chan. (1999). KDD Cup 1999 Data. UCI Machine Learning Repository. https://doi.org/10.24432/C51C7N.

Yilmaz, Ibrahim, Rahat Masum, and Ambareen Siraj. (2020). "Addressing Imbalanced Data Problem with Generative Adversarial Network For Intrusion Detection." In 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI). 25–30. https://doi.org/10.1109/IRI49571.2020.00012

Zavrak, Sultan, and Murat İskefiyeli. (2020). "Anomaly-Based Intrusion Detection From Network Flow Features Using Variational Autoencoder." IEEE Access 8 (2020), 108346-108358. https://doi.org/10.1109/ACCESS.2020.3001350

Zingo, Pasquale, and Andrew Novocin. (2021). “Introducing the TSTR Metric to Improve Network Traffic GANs.” In: Arai, K. (eds) Advances in Information and Communication. FICC 2021. Advances in Intelligent Systems and Computing, vol 1363. Springer, Cham. https://doi.org/10.1007/978-3-030-73100-7_46