https://doi.org/10.5281/zenodo.14054137

A Review on Machine Learning Based Intrusion Detection System: Techniques, Public Datasets and Challenges

Goce Stevanoski[1], Marko Porjazoski[2], Ivo Paunovski[3], Monika Kachurova[1], Aleksandar Risteski[2]
[1]Military Academy General Mihailo Apostolski, Goce Delcev University, Skopje, N. Macedonia
{goce.stevanoski, monika.kacurova}@ugd.edu.mk
[2]Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University, Skopje, N. Macedonia, {marko, acerist}@feit.ukim.edu.mk
[3]University American College Skopje, N. Macedonia, ivo.paunovski@uacs.edu.mk

Dedicated to Prof. Mile J. Stankovski on his anniversary

***Abstract*—This review paper focuses on the machine learning techniques used by the research community for detecting anomalies in network traffic in order to show intrusion activities. The paper states some of the state-of-the-art machine learning techniques used in anomaly detection. Furthermore, the paper briefly describes the evolution of training datasets, highlights the main representatives, and points out some of the shortcomings that should be taken into consideration while addressing the issues of anomaly detection with intrusion detection systems in modern IT network environments. The review emphasizes the perspectives for future work on this subject by presenting the challenges of training machine learning models for anomaly detection in intrusion detection systems.**

Keywords—key anomaly detection, intrusion detection systems, machine learning, IDS datasets

I. INTRODUCTION

Detecting anomalies in network traffic is a method that can significantly improve network security[1]. The intrusion detection system's (IDS's) ability to distinguish between the normal (benign) traffic and abnormal (malicious) traffic increases the confidence in its deployment and helps in mitigating the operational malfunctions and security vulnerabilities.

The research community continuously searches for ways to ensure efficient security and quality of service in large-scale networks. The growth of new communication technologies and services, coupled with the rising number of web users, applications, interconnected network devices, and services, makes computer networks large and increasingly complex.

The term "anomaly" has various interpretations. Barnett and Lewis define an anomaly in a data set as an "observation (or a subset of observations) which appears to be inconsistent with the remainder of that set of data" [2]. Chandola et al. describe it as "patterns in data not conforming to a well-defined notion of normal behavior"[3]. According to Lakhina et al., "anomalies are unusual and significant changes in a network's traffic levels, which can often span multiple links"[4]. Hoque et al. define it as "non-conforming interesting patterns compared to the well-defined notion of normal behavior". These definitions highlight the role of anomalies in generated datasets and the importance of understanding the concept of normality in developing solutions to detect network anomalies.

Although it may seem straightforward, defining a region that denotes normal behavior and identifying anomalies that deviate from this pattern is quite challenging. The primary goals in this field are faster diagnosis, reduced complexity, and appropriate correction of causes.

This ability of machine learning (ML) models to detect anomalies by distinguishing normal from abnormal, can be included in intrusion detection systems (IDS) and enhance the security of IT networks. Achieving precise anomaly detection in IDS requires that a solid baseline of normal data points, structured in a dataset, must be established. Datasets consist of different features and measurements created by collecting data points from external activities and internal activities of the network hosts [5]. Data points collected from external activities represent the packet flows, while the internal activities are represented by system logs, host statistics etc., generated from the various processes and services running on the hosts. In the field of IT network security, ML models for anomaly detection can be trained with two different approaches: first by using data points from one data source and second by using data points from a combination of data sources.

The potential impact of this research effort is immense, it should inspire development of numerous models, algorithms, and mechanisms to ensure the security and health of increasingly large and complex network systems. As in [6],[7], the focus is primarily on the most popular techniques but also this paper is reviewing other important topics such as datasets, challenges and recommendations.

II. Related Work

In this part, we are going to review and synthesize existing research on anomaly detection in intrusion detection systems. The last 20 years have been clearly marked by many proposals for ML implementation in anomaly detection. This was stated by the authors of [8], where they emphasize that "ML is suitable for implementing IDS solutions in real time with no (or little) human supervision." This means that machine learning IDSs, by detecting anomalies in generated IT network data, can effectively detect known and unknown intrusions. IDS powered by machine learning algorithms can deliver robust protective measures against network intrusions and enable swift detection of security vulnerabilities, allowing for swift corrective actions to minimize potential harm.

The authors of [9] conducted a systematic review of data mining in big data for intrusion detection and concluded that the most used data mining techniques are Decision Tree (DT), Bayesian Networks (BN), k-Means, Artificial Neural Network (ANN), and Support Vector Machine (SVM). In the work of [10], the researchers presented various research papers that show the importance of ML in intrusion detection systems, particularly by implementing different classifiers for intrusion detection. The authors conclude that ensemble and hybrid classifiers have shown better performance than single classifier models. This is similar to the results of the work in [11], where the authors reviewed various studies and concluded that ensemble and hybrid classifiers demonstrated the highest predictive accuracy and detection rate, outperforming their single-classifier counterparts.

The authors of [12] focus on studying big data classification for intrusion detection. The approach with big data classification can help optimize the process of intrusion detection. Their work led them to the conclusion that the techniques and technologies used should be expanded toward verdict solutions to overcome big data classification difficulties. Some of the optimization techniques were addressed in [13], where the authors conducted a study that reviewed several optimization algorithms proposed in the past decade, providing researchers with insights into various intrusion detection approaches and highlighting the limitations and opportunities for designing improved algorithms. Each algorithm has its own strengths and weaknesses, which often need to be combined with other techniques for better performance.

The pattern matching relevant to network intrusion detection is addressed in this 2023 review [14], where the authors have outlined the widely adopted research and methods. The review assessed advancements in signature-based systems' pattern matching techniques, using both fixed strings and regular expressions, for handling large-data volumes in line with modern trends. The review discusses how machine learning techniques are increasingly important for analyzing malicious network traffic and strengthening defense against network intrusions, with the need to correlate large volumes of data from multiple network endpoints and security applications collected and analyzed by Security Event and Incident Management (SIEM) systems.

From the perspective of training ML models with data points collected from external host activities, numerous review papers were published specifically on working with packet flow data points. The authors of [15] have been working on synthesizing broadly published research papers regarding datasets in the field of Intrusion Detection Systems used for performance evaluation of ML-based IDS. The authors conclude that there is a need for updating contemporary datasets since they do not reflect new trends of attacks that include different processes and technologies. Additionally, the patterns in the datasets don't reflect realistic network scenarios. Similar results were presented in [16], where a study on various techniques for intrusion detection was conducted and shows that there might be a problem with updating the information for new attacks, resulting in high false alarms or poor accuracy. The study reviewed popular datasets, stating that there should be more new datasets with broader malware activities. In 2022, the authors of [17] used four popular datasets—KDD'99, NSL-KDD, CICIDS2017, and UNSW-NB15—to create a summarization and offer taxonomy. They state that feature selection is critical and almost always required for improving performance. Additionally, they address the issue of data imbalance as a problem that can be overcome with a sampling approach. Regarding the process of selecting dataset features for better results, the authors of [18] demonstrated the use of Feature Elimination and Principal Component Analysis (PCA) over various classification algorithms such as Decision Tree, Naïve Bayes, and Support Vector Machine, with the Decision Tree showing the best results. Working on a similar issue, the authors of [19] provided a comprehensive overview of various ML techniques for IDS, concluding that 58.33% of the research papers include the KDD-NSL dataset. The authors suggest using more modern datasets in future research and note that ensemble and hybrid classifiers have outperformed their single-classifier equivalents.

Although most of the work is done using datasets with external data points, there are studies that utilize data points from internal host activities [20], [21], [22]. However, this is not reflected in many review studies, and the overall approach is missing. Some recent reviews address the importance of combining external and internal data points in the so-called Predictive IDS approach [23]. Additionally, the authors of [24] focused on reviewing deep learning techniques for IDS by analyzing data points from network flow and system logs, although they do not emphasize the importance of datasets in developing IDS. One of the most comprehensive reviews was done in 2023 by the authors of [25], where they present internal data points as part of modern security architectures and devices responsible for securing the IT network perimeter.

The authors of [26], besides providing an extended overview of the taxonomy of network attacks, their tools, the most relevant features to detect them, data sources of IDSs, data types in datasets, the architecture of the IDSs, and the detection modes, are also addressing the challenges for future investigation. The authors point out that future challenges will be in defining what constitutes an anomaly and what does not. They also emphasize the tasks required for preprocessing data, as the datasets are enormous, and proper feature selection represents a significant challenge.

III. Machine Learning Techniques For IDS

As stated in the previous section, ML algorithms are widely used in conducting research studies on anomaly detection for IDS. These research studies have defined several state-of-the-art methods that can be used as benchmarks for future reference. Broadly, ML algorithms can be divided into two categories: Supervised ML and Unsupervised ML. The former uses labeled data for ML model development, while the latter uses unlabeled data.

The development of ML models includes the use of algorithms like k-Nearest Neighbor (kNN), Decision Tree (DT), Naïve Bayes (NB), and Support Vector Machine (SVM) for supervised ML, and K-means and Self-Organizing Map (SOM) for unsupervised ML model development. The application of all the mentioned ML algorithms in the field of anomaly detection, even outside of IDS in IT networks, defines them as representatives of state-of-the-art ML algorithms.

IV. Overview Of The Public Datasets

Anomaly detection for IDS is an important and hot topic for many researchers in the field of cybersecurity. They implement it in all aspects of IT network security, especially when processing security-related data to detect malicious activities. This requires quality IT network datasets that can help in setting the baseline for developing new systems and applications and setting benchmarks for future system evaluation.

Fortunately, the research community continuously develops network-related datasets that can be used for conducting research in the field of anomaly detection in IT network traffic. Although these datasets are not universally applicable to all systems, they can help in defining the core aspects of anomaly detection in intrusion detection systems.

In the next sections, we will present some of the most common publicly available network datasets. We will review two representatives of the datasets in three distinctive periods of development, which we have named: Initial Dataset Collections, Progressive Dataset Collections, and Advanced Dataset Collections. This selection of datasets reflects the historical aspect of dataset development, their usage by the research community, the diversity of attacks covered by the dataset, and the volume and quality of data.

A. Initial Dataset Collections

This collection of datasets was built in the period from the mid-90s until the mid-2000s and represents IT network datasets that are missing records of data extracted from modern network protocols. Additionally, these datasets usually contain inconsistent and duplicate data points in the records[27]. Some of the representatives include KDD Cup 1999 and NSL-KDD Dataset, which are described in the following paragraphs.

- KDD Cup 1999 Dataset

Developed by DARPA, KDD CUP 99 is a dataset that doesn't belong to packet capture or flow capture formats. This dataset consists of features collected from TCP connections and failed logins. In this dataset, there is no IP address. With more than 5 million data points, this dataset includes 20 different types of attacks. The KDD CUP 99 has been broadly used over the last 25 years by the research community in building and evaluating systems for anomaly detection. Since it was created in 1998, its biggest disadvantage is that the protocols used are old and don't reflect the current networking environment[27].

- NSL-KDD Dataset

NSL-KDD supersedes the KDD CUP 99. Various implementations and reviews have shown that KDD CUP 99 has a problem with a large number of duplicates in the dataset. This led to the development of NSL-KDD. This dataset contains 150,000 data points and is divided into training and testing subsets designed for intrusion detection. In terms of similarity to KDD CUP 99, NSL-KDD has the same features, and the traffic is dated from 1998. This means that the network traffic represented in this dataset is old and doesn't represent new IT network environments [28].

B. Progressive Dataset Collections

This collection includes datasets developed from the early 2010s to the mid-2010s. The need for the development of this collection arose from the disadvantages seen in datasets like KDD Cup 1999 and NSL-KDD. Some of the datasets from this period are shown in the following paragraphs.

- CTU-13 Dataset

Developed in 2011 by CTU University, Czech Republic, the CTU-13 dataset encompasses IT network traffic data points in three formats: packet, unidirectional flow, and bidirectional flow. This dataset represents network data from botnet attacks in 13 different scenarios. The traffic is labeled in three ways: one set is labeled as benign network traffic, another set is labeled as malicious network traffic, and the rest is labeled as background network traffic. This means that the background traffic can be either malicious or benign, and it is up to the user to divide the dataset into training and testing subsets.

- UNSW-NB15 Dataset

Created in 2015, the UNSW-NB15 Dataset represents IT network traffic captured in packet-based format. This dataset contains malicious and benign network traffic generated with the IXIA Perfect Storm tool over a period of 31 hours. Malicious traffic includes 9 different families of attacks such as DoS, exploits, fuzzers, backdoor, etc. The dataset can also be used in flow-based format, and it has additional features included. UNSW-NB15 is divided into two subsets: one for training and one for testing purposes. In comparison to the KDD datasets, this UNSW-NB15 has 45 distinct IP addresses.

C. Advanced Dataset Collections

The advanced datasets collection represents datasets with the most sophisticated collection of data points. Here, the research community has invested heavily in developing these datasets, and they are built with a diverse range of IT network protocols at various levels of the OSI model. Additionally, the datasets include records of data points representing various attacks, developed following different attack methodologies. Some of the representatives are shown in the following paragraphs.

- CICIDS 2017

One of the most advanced datasets for training intrusion detection systems is CICIDS 2017. Developed by the Canadian Institute for Cybersecurity, this dataset encompasses network traffic in bidirectional flow-based format and packet-based format. It consists of more than 80 network traffic features, and the researchers have extracted additional metadata about the attacks in the dataset and the IP addresses used. The dataset includes data points collected for normal user behavior generated with scripts. CICIDS 2017 contains various attacks such as Heartbleed, DoS, DDoS, SSH brute force, web attacks, and infiltration attacks.

- CUPID

With more than 50GB of data, the CUPID dataset represents one of the most robust datasets developed for training and evaluating IDS systems. This dataset has 179 unique hosts and generated network traffic that contains 75% TCP-based and 7.5% UDP-based protocols. This dataset includes network traffic that uses Internet Hypertext Transfer Protocol (HTTP), Simple Mail Transfer Protocol (SMTP), Internet Control Message Protocol (ICMP), Lightweight Directory Access Protocol (LDAP), Kerberos, Distributed Computing Environment (DCE) / Remote Procedure Call (RPC), Server Message Block (SMB), and Network Basic Input/Output System (NetBIOS). The rich variety of protocols included in this dataset is possible because of the different services involved in the traffic generation process, such as email, DNS lookups, or active directory accesses. Additionally, the dataset was created by considering the attacker's perspective and following the stages within the cyber kill chain attack strategy.

The presented datasets are just a small sample of the developed datasets used by the research community in developing ML models for anomaly detection in IDS. A more in-depth analysis of the various data points and dataset features was done by the authors of [29], where they analyze the strengths, weaknesses, limitations, and use concerns of different IDS datasets in their 2024 review paper. Their work makes it clear that each dataset has distinct attributes and compromises. Additionally, they suggest creating a unified approach to building datasets that encompass different dataset features and data points collected from different data sources. This is particularly interesting since, out of the 37 analyzed datasets, only two, MACCDC2012 and CICIDS-CRIME2018, consider data points from internal hosts like system logs and events. This means that a large number of research efforts in anomaly detection for IDS have been conducted without considering the information that can be extracted from internal data points.

Overall, the datasets used for anomaly detection in intrusion detection systems are focused on parameters collected or extracted from IT network traffic. This is clearly represented by the authors in [30], who collected data about 34 datasets used for training IDS in anomaly detection. By examining their findings, we can conclude that, although there are datasets with various approaches and collected features, these datasets lack syslog data points, which can be crucial in detecting anomalies in the network.

V. FEATURE CHALLENGES

When we take into consideration already proven ML algorithms for anomaly detection in IDS, we can conclude that the algorithms mentioned in the second part of these papers are some of the state-of-the-art ML algorithms for detecting anomalies in IT networks. This means they can determine intrusion with a very high probability. However, this doesn't come without some considerations that need to be addressed. These considerations should mostly apply to the datasets used for the training and testing of ML models.

A. Unified Dataset

Current datasets used by the research community for developing IDS have unique qualities that should be thoroughly evaluated before use[29]. This evaluation should include an analysis of the attack types, the accuracy of the network traffic patterns, the size and complexity of the data, and whether the data is labeled or not. All of this should be done to ensure the selected dataset is appropriate for conducting the research. This means the process of selecting appropriate datasets is inefficient since researchers must spend more time evaluating the dataset. Dedicating one unified dataset for broader, or possibly all, future developments of IDS would optimize the process of developing anomaly detection systems based on ML.

B. Future Evolution

The continuous advancement of technology produces new and unforeseen threats to IT network environments. This is a challenging aspect for the future development of IDS based on anomaly detection. These systems, built on data points collected from previous periods, will not have the new threat patterns incorporated into the ML models. Therefore, a mechanism for continuous monitoring and updating should be introduced to help ML models evolve. How this will be introduced without affecting the current running process is one of the tasks for future research.

C. Multi-Data Source Approach

Since IT network environments today are complex structures with various security devices, IDS should be able to follow the baseline structure of the IT network, including all generated security data points, regardless of the data source from which they are extracted. In that sense, a multi-data source approach should be established. Data points should be extracted from both external and internal sources and encompassed in a reliable dataset that can be used to train ML models for anomaly detection. In the previous section, we noted that a recent study evaluated 37 datasets, but only two of them had data points extracted from internal data sources. This is a very low number of datasets that can offer a multi-data source approach, limiting the possibilities and objectives for future research.

D. Adversarial Attacks and Data Poisoning

Adversarial attacks, which involve the manipulation of input data to evade detection, and data poisoning, where attackers inject malicious samples into the training data, represent significant issues in developing ML models for anomaly detection with IDS[31]. These attacks can compromise the

integrity and effectiveness of ML models. Prospective research should develop techniques for creating new generations of datasets resilient to these types of attacks and mitigate the effects of adversarial attacks and data poisoning. The approach should involve adversarial training, data sanitization, and robust learning algorithms[33][34].

E. Imbalance in Dataset Classes

The balance between benign and malicious data points in a dataset influences the performance of ML models. If there is an imbalance between these two classes, the model's performance will suffer significantly[35]. This can lead to biased models that perform poorly in detecting anomalies. Efforts to overcome these issues should include developing advanced techniques to address class imbalances. These techniques should involve building robust models with improved sampling methods, data augmentation, and the utilization of synthetic data [36][37].

F. Data Anonymization

Data anonymization is a process that can help secure personal and sensitive information embedded in the dataset by incorporating various modifications and exclusions, which leads to developing ML models with datasets that have modified or excluded private and sensitive information. This is an important issue since the manipulation of significant information in the dataset can lead to an irrelevant dataset. Irrelevance comes from the loss of essential features necessary for detecting intrusions[38][39]. The challenge is to preserve the performance of the model while anonymizing private or sensitive data. This should be done by introducing advanced techniques[40] for data anonymization while preserving the utility of the data and restricting access to private or sensitive information.

VI. CONCLUSION

In the previous sections, it was stated that detecting anomalies in IT network data can significantly help in denying and mitigating potential security and operational threats. The state-of-the-art ML algorithms for anomaly detection are available and can support us in developing ML models with reliable and efficient performance. Supervised ML algorithms have performed much better than unsupervised ML algorithms. Evaluations on advanced datasets have shown that the performance of the models is highly dependent on the type of ML algorithm and the dataset used in the training and testing phases.

In this paper, we presented that, although some algorithms have shown remarkable results, there are still challenges that need to be addressed when discussing intrusion detection in IT networks. In the future, the research community can significantly improve intrusion detection by addressing these challenges, which are mostly identified in the domain of datasets used for training and testing ML models. The diversity of information that datasets can hold by collecting data points from the network makes this area exciting, promising, and challenging for investigation. Making a breakthrough by resolving the presented challenges will foster the development of new generations of IDS based on detecting anomalies in datasets created from internal and external data sources.

REFERENCES

[1] W. A. Ali, K. N. Manasa, M. Bendechache, M. Fadhel Aljunaid, and P. Sandhya, "A review of current machine learning approaches for anomaly detection in network traffic," Journal of Telecommunications and the Digital Economy, vol. 8, no. 4, pp. 64–95, 2020.

[2] V. Barnett and T. Lewis, Outliers in statistical data, vol. 3, no. 1. Wiley New York, 1994.

[3] V. Chandola, Anomaly detection for symbolic sequences and time series data. University of Minnesota, 2009.

[4] A. Lakhina, M. Crovella, and C. Diot, "Characterization of network-wide anomalies in traffic flows," in Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, 2004, pp. 201–206.

[5] T. Guemmah, H. El Fadili, and S. Hraoui, "A Review and Synthesis for Framing the Use of Artificial Intelligence in Cybersecurity," in 2023 7th IEEE Congress on Information Science and Technology (CiSt), IEEE, Dec. 2023, pp. 44–49. doi: 10.1109/CiSt56084.2023.10409914.

[6] M. Thottan, G. Liu, and C. Ji, "Anomaly detection approaches for communication networks," Algorithms for next generation networks, pp. 239–261, 2010.

[7] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," Computer networks, vol. 51, no. 12, pp. 3448–3470, 2007.

[8] H. Aludhilu and R. R. Puente, "A systematic literature review on intrusion detection approaches," Revista Cubana de Ciencias Informáticas, vol. 14, no. 1, pp. 58–78, 2020.

[9] F. Salo, M. Injadat, A. B. Nassif, and A. Essex, "Data mining with big data in intrusion detection systems: A systematic literature review," arXiv preprint arXiv:2005.12267, 2020.

[10] U. S. Musa, M. Chhabra, A. Ali, and M. Kaur, "Intrusion Detection System using Machine Learning Techniques: A Review," in 2020 International Conference on Smart Electronics and Communication (ICOSEC), IEEE, Sep. 2020, pp. 149–155. doi: 10.1109/ICOSEC49089.2020.9215333.

[11] O. Ambavkar, P. Bharti, A. Chaurasiya, R. Chauhan, and M. Palinje, "Review on IDS based on ML Algorithms," Int J Res Appl Sci Eng Technol, vol. 10, no. 11, pp. 169–174, Nov. 2022, doi: 10.22214/ijraset.2022.47284.

[12] N. Mishra and S. Mishra, "A Review on Big Data Classification: using Machine Learning Technique to Classify Intrusion," IJRAR-International Journal of Research and Analytical Reviews (IJRAR), vol. 7, no. 1, pp. 162–165, 2020.

[13] S. Sadiq and A. S. Eesa, "OPTIMIZATION ALGORITHMS FOR INTRUSION DETECTION SYSTEM: A REVIEW," International Journal of Research -GRANTHAALAYAH, vol. 8, no. 8, pp. 217–225, Aug. 2020, doi: 10.29121/granthaalayah.v8.i8.2020.1031.

[14] N. Venkatesh and Prof. P. Kashikar, "A Literary Review of Pattern Matching Techniques in Network Intrusion Detection," International Journal of Current Science Research and Review, vol. 06, no. 09, Sep. 2023, doi: 10.47191/ijcsrr/V6-i9-40.

[15] A. Thakkar and R. Lohiya, "A Review of the Advancement in Intrusion Detection Datasets," Procedia Comput Sci, vol. 167, pp. 636–645, 2020, doi: 10.1016/j.procs.2020.03.330.

[16] S. Patidar, P. Parihar, and C. Agrawal, "A Review of Intrusion Detection Datasets and Techniques," SMART MOVES JOURNAL IJOSCIENCE, vol. 6, no. 3, pp. 14–22, Mar. 2020, doi: 10.24113/ijoscience.v6i3.277.

[17] E. E. Abdallah and A. F. Otoom, "Intrusion detection systems using supervised machine learning techniques: a survey," Procedia Comput Sci, vol. 201, pp. 205–212, 2022.

[18] S. Singh and S. Banerjee, "Machine Learning Mechanisms for Network Anomaly Detection System: A Review," in 2020 International Conference on Communication and Signal Processing (ICCSP), IEEE, Jul. 2020, pp. 0976–0980. doi: 10.1109/ICCSP48568.2020.9182197.

[19] U. S. Musa, S. Chakraborty, M. M. Abdullahi, and T. Maini, "A Review on Intrusion Detection System using Machine Learning Techniques," in 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), IEEE, Feb. 2021, pp. 541–549. doi: 10.1109/ICCCIS51004.2021.9397121.

[20] R. Pokhrel, P. Pokharel, and A. Kumar Timalsina, “Anomaly-Based – Intrusion Detection System using User Profile Generated from System Logs,” International Journal of Scientific and Research Publications (IJSRP), vol. 9, no. 2, p. p8631, Feb. 2019, doi: 10.29322/IJSRP.9.02.2019.p8631.

[21] Lu et al., “Integrating Traffics with Network Device Logs for Anomaly Detection,” Security and Communication Networks, vol. 2019, pp. 1–10, Jun. 2019, doi: 10.1155/2019/5695021.

[22] Y.-D. Lin, Z.-Y. Wang, P.-C. Lin, V.-L. Nguyen, R.-H. Hwang, and Y.-C. Lai, “Multi-datasource machine learning in intrusion detection: Packet flows, system logs and host statistics,” Journal of Information Security and Applications, vol. 68, p. 103248, Aug. 2022, doi: 10.1016/j.jisa.2022.103248.

[23] P. H. Durole and M. Agarwal, “‘A Comprehensive Review of Advanced Artificial Intelligence Techniques to Enhance Intrusion Detection Systems,’” in 2024 IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS), IEEE, Feb. 2024, pp. 1–7. doi: 10.1109/SCEECS61402.2024.10482305.

[24] Y. R. Gumma and S. Peram, “Review of Cybercrime Detection Approaches using Machine Learning and Deep Learning Techniques,” in 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), IEEE, Jun. 2024, pp. 772–779. doi: 10.1109/ICAAIC60222.2024.10575058.

[25] J. L. Gutierrez-Garcia, E. Sanchez-DelaCruz, and M. del P. Pozos-Parra, “A Review of Intrusion Detection Systems Using Machine Learning: Attacks, Algorithms and Challenges,” 2023, pp. 59–78. doi: 10.1007/978-3-031-28073-3_5.

[26] S. Hajj, R. El Sibai, J. Bou Abdo, J. Demerjian, A. Makhoul, and C. Guyeux, “Anomaly-based intrusion detection systems: The requirements, methods, measurements, and datasets,” Transactions on Emerging Telecommunications Technologies, vol. 32, no. 4, Apr. 2021, doi: 10.1002/ett.4240

[27] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the KDD CUP 99 data set,” in 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, IEEE, Jul. 2009, pp. 1–6. doi: 10.1109/CISDA.2009.5356528.

[28] R. Bala, “A REVIEW ON KDD CUP99 AND NSL-KDD DATASET,” International Journal of Advanced Research in Computer Science, vol. 10, no. 2, pp. 64–67, Apr. 2019, doi: 10.26483/ijarcs.v10i2.6395.

[29] A. Khanan, Y. Abdelgadir Mohamed, A. H. H. M. Mohamed, and M. Bashir, “From Bytes to Insights: A Systematic Literature Review on Unraveling IDS Datasets for Enhanced Cybersecurity Understanding,” IEEE Access, vol. 12, pp. 59289–59317, 2024, doi: 10.1109/ACCESS.2024.3392338.

[30] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, “A survey of network-based intrusion detection data sets,” Comput Secur, vol. 86, pp. 147–167, Sep. 2019, doi: 10.1016/j.cose.2019.06.005.

[31] J. Lin, L. Dang, M. Rahouti, and K. Xiong, “ML Attack Models: Adversarial Attacks and Data Poisoning Attacks,” Dec. 2021.

[32] P. Tavllali and V. Behzadan, “Short Review on Supervised Learning Under Adversarial Label Poisoning and Clean-Label Data Poisoning,” in 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE), IEEE, Jul. 2023, pp. 378–382. doi: 10.1109/CSCE60160.2023.00067.

[33] T. Chaalan, S. Pang, J. Kamruzzaman, I. Gondal, and X. Zhang, “The Path to Defence: A Roadmap to Characterising Data Poisoning Attacks on Victim Models,” ACM Comput Surv, vol. 56, no. 7, pp. 1–39, Jul. 2024, doi: 10.1145/3627536.

[34] L. Dang, T. Hapuarachchi, K. Xiong, and J. Lin, “Improving Machine Learning Robustness via Adversarial Training,” in 2023 32nd International Conference on Computer Communications and Networks (ICCCN), IEEE, Jul. 2023, pp. 1–10. doi: 10.1109/ICCCN58024.2023.10230138.

[35] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, “Classification with class imbalance problem,” Int. J. Advance Soft Compu. Appl, vol. 5, no. 3, pp. 176–204, 2013.

[36] S. Bagui and K. Li, “Resampling imbalanced data for network intrusion detection datasets,” J Big Data, vol. 8, no. 1, p. 6, Dec. 2021, doi: 10.1186/s40537-020-00390-x.

[37] R. Panigrahi et al., “A Consolidated Decision Tree-Based Intrusion Detection System for Binary and Multiclass Imbalanced Datasets,” Mathematics, vol. 9, no. 7, p. 751, Mar. 2021, doi: 10.3390/math9070751.

[38] B. Raghunathan, The complete book of data anonymization: from planning to implementation. CRC Press, 2013.

[39] S. Murthy, A. Abu Bakar, F. Abdul Rahim, and R. Ramli, “A Comparative Study of Data Anonymization Techniques,” in 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), IEEE, May 2019, pp. 306–309. doi: 10.1109/BigDataSecurity-HPSC-IDS.2019.00063.

[40] J. F. Marques and J. Bernardino, “Analysis of Data Anonymization Techniques.,” in KEOD, 2020, pp. 235–241