IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

Cyber Attack Prediction: From Traditional Machine Learning to Generative Artificial Intelligence

Shilpa Ankalaki[1], Aparna Rajesh A[2], Pallavi M[3], Geetabai S Hukkeri[1], Tony Jan[4], Ganesh R. Naik[4,5,6]

[1]Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, Karnataka 576104, India
[2]Department of CSE, SoET, Centurion University of Technology and Management Bhubaneswar, Odisha.
[3]Department of School of Computer and Science and Engineering, Presidency University, Bangalore
[4]Centre for Artificial Intelligence Research and Optimization (AIRO), Design and Creative Technology Vertical, Torrens University, Ultimo, NSW 2007, Australia
[5]Design and Creative Technology Vertical, Torrens University, Wakefield Street, Adelaide SA 5000
[6]College of Medicine and Public Health, Flinders University, Adelaide, Australia

Corresponding author: Shilpa Ankalaki (e-mail: shilpa.ankalaki@manipal.edu).

ABSTRACT The escalating sophistication of cyber threats poses significant risks to individuals, organizations, and nations. Cybercrime, encompassing activities like hacking and data breaches, has severe economic and societal consequences. In today's interconnected world, robust cybersecurity measures are paramount to mitigate these risks and protect sensitive information. However, traditional security solutions struggle to keep pace with the evolving threat landscape. Artificial Intelligence (AI) offers a powerful arsenal of techniques to address these challenges. This paper explores the application of AI methods, including Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), Explainable AI (XAI), and Generative AI, in solving various cybersecurity problems. This paper presents a comprehensive analysis of AI techniques for enhancing cybersecurity. Key contributions include: i) A comparative study of ML and DL methods: Evaluating their accuracy, applicability, and suitability for various cybersecurity challenges. ii) An investigation into XAI approaches: Enhancing the transparency and interpretability of AI-powered security solutions, particularly in anomaly detection. iii) An exploration of emerging trends in Generative AI (Gen-AI) and NLP: Examining their potential to simulate and mitigate cyber threats through advanced techniques like threat intelligence generation and attack simulations. iv)Application of GenAI in cybersecurity and real-world products of GenAI for cyber security. This research aims to advance the state-of-the-art in AI-driven cybersecurity by providing insights into effective and reliable solutions for mitigating cyber risks and improving the overall security posture.

INDEX TERMS Cybersecurity, cyber-attack prediction, Machine Learning, Deep Learning, Explainable AI, and Generative AI

I. INTRODUCTION

With rapid technological advancements and increasing interconnectivity in our community, the significance of security solutions and measures for mitigation will be more essential. Technological advancements make everyone's life easier and more convenient in all aspects but concurrently present several challenges. One of the significant challenges is the swift increase in cybersecurity threats alongside technological advancements. As technological progress advances and businesses increasingly rely on digital platforms, the spectrum of cyber-attacks has grown more ominous. Such attacks have the potential to inflict severe damage on individuals and organizations alike, leading to financial setbacks, tarnished reputations, and even jeopardizing national security. Therefore, it is imperative for governments, businesses, and individuals to accord the highest priority to cybersecurity measures to safeguard their respective interests [1]. Considering all these aspects, cybersecurity has become significantly more important for researchers and professionals. It encompasses a wide array of elements, including tools, techniques, policies, security measures, guidelines, risk-mitigation strategies, training, best practices, and innovative technologies. These components collectively aim to protect cyberspace and user assets [2]. Cybersecurity refers to mechanisms that protect

systems against threats and vulnerabilities to ensure the efficient delivery of accurate services to users. Owing to the rapid increase in data volume, ensuring security has become a major challenge in cybersecurity. Modern hackers have profound knowledge of systems and programming expertise, allowing them to exploit well-protected hosts. Some attacks with immense destructiveness in the last few years are listed below.

- In May 2021, the Colonial Pipeline, a major supplier of gasoline to the eastern United States, fell victim to a ransomware attack, resulting in the shutdown of its pipeline for an extended duration. The attack was orchestrated by a Russian hacking group known as DarkSide, which demanded a $4.4 million ransom payment in Bitcoin. These cyberattacks triggered widespread panic and fuel shortages across numerous states [3].
- SolarWinds Supply Chain Attack: It was discovered in December 2020 that the Orion network monitoring software had been compromised and malicious malware had been introduced into SolarWinds software. Numerous government institutions and commercial businesses have been affected by the breach [4].
- Log4J Vulnerability: This zero-day attack, known as Log4Shell before an official CVE designation, was assigned to the security industry in late 2021. 100s of Millions of devices have been affected [5].
- The WannaCry ransomware assault occurred in May 2017 and affected over 200,000 systems across 150 countries. The compromised PC files were encrypted by ransomware, which then requested payment in Bitcoin to unlock them. The attack resulted in extensive disturbances, encompassing the shutdown of multiple hospitals in England [6].

As is evident from numerous studies, cybercrime has harmed several organizations, companies, and people in recent years. As cyber threats evolve in complexity and frequency, conventional cybersecurity measures are proving to be insufficient to detect and counter emerging attack methodologies [7]. Cyberattack defense for computer-based systems has become increasingly difficult. It is necessary to design more effective and efficient cybersecurity solutions to prevent cyberattacks.

To mitigate security risks and minimize their consequences, the cybersecurity sector has directed its research and development endeavors toward specific focal points. It is widely acknowledged within the cybersecurity community that the complete elimination of cyber threats is unattainable. Consequently, the predominant strategies tend to be reactive rather than proactive. Notably, in recent years, considerable scholarly attention has been paid to incident response and intrusion detection, yielding promising results. Nonetheless, these efforts primarily address post-event scenarios, limiting their effectiveness in pre-emptive measures [8]. Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) are increasingly acknowledged as potent instruments for tackling cyber security challenges. They possess the capacity to augment the functionalities of current cybersecurity systems and identify unrecognized threats [9]. AI, ML, and DL are frequently used interchangeably. Figure 1 illustrates the interconnectedness among AI, ML, and DL. AI serves as a broad domain akin to the universe, whereas ML operates within the realm of AI as a subset, and DL further specializes as a subset within ML. AI provides the ability to sense, reason, act, and adapt. ML is an application of AI that enables machines to learn automatically and improve their past data experience. DL is an application of ML that utilizes complex algorithms and deep neurons to train a model. This requires a large amount of data.

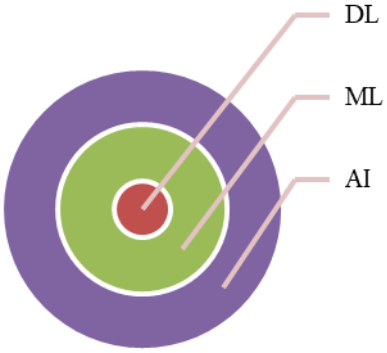

FIGURE 1. Relation between AI, ML, and DL.

A. How AI Enhances Cybersecurity: Key Benefits and Application

ML algorithms are trained with historical experience, to predict future outcomes in a way that resembles human decisions. ML algorithms are widely used in cybersecurity for identifying security threats and breaches, is an example of an ML application. In the past few years, automated security tools based on ML have been created to provide an autonomous response to the threats by using clustering, classification and regression techniques [10]. Proactive vulnerability management is also where AI & ML are used. AI/ML based tools like User and Entity Behavior Analytics (UEBA) work on the principle that malware is often detected by monitoring user interactions on servers and service endpoints, helping identify such unusual behavior. This method allows organisations to identify and mitigate risks proactively, frequently before exploits are made public0 or patched [11]. AI finds application in diverse areas, from very simplistic recurring processes to more advanced applications of AI like cybersecurity, where AI counteracts advanced cyber threats. This is a new technology that is transforming machines into machines that can also think, thus making more human-like decisions, performing tasks and automating them using assisted, augmented and autonomous intelligence.

However, in today's digital ecosystem, where cyber risks are continuously evolving, traditional security measures often lack the agility and sophistication needed to effectively defend against 21st-century cyberattacks [12], such as zero-day vulnerabilities. AI in cybersecurity can help organizations to make smarter decisions, detect Network Invasions, and heal the effects of cyberattacks. DL is a subbranch of machine learning that focuses on learning the representation by passing the information through multiple layers of transformations which makes it more accurate for classification and regression tasks. The proposed Defense framework DL-based solution can be increasingly used for defense purposes in Cybersecurity, as DL-based defense mechanisms are already in place in different combinations to automate detection of cyber threats, such systems getting trained progressively and improving as time passes [13]. The adoption of AI in cybersecurity comes with its own set of challenges. To function properly, AI systems require large amounts of data, which means processing can consume a lot of resources. Moreover, false alarm complexity might damage the user trust [20], and late threat responses could make the entire system less efficient. AI-based security systems can be vulnerable to cyber-attacks that target the system itself. Nevertheless, ongoing research is improving the resilience of AI against such cyber threats[14].

B. AI Limitations That Highlight the Need for XAI in Cybersecurity

AI algorithms can also be susceptible to adversarial attacks, where attackers manipulate input data to trick the AI system before the attack takes place, highlighting the need for stringent testing and evaluation. There are several key issues, one amongst these is that of using AI systems by malicious actors or using AI systems as vectors for attacks. Evasion attacks, for example, can allow attackers to modify malware files to be mistaken as benign files to detection systems that rely on machine learning to detect malicious files. Apart from the aforementioned threats, AI-enabled cybersecurity systems are also susceptible to a wide range of other threats, including communication interception, service failures, accidents, environmental disasters, legal issues, and other security threats, power outages, and other physical damage, all of which might cause the malfunctioning of these systems [15].
AI is the fundamental technology of Industry 4.0, and it also plays a significant role in advancing cyber security services and management [16] . Various AI techniques, especially the ML and DL algorithms, have been utilized for malware detection, anomaly detection, and network traffic analysis. ML approaches have been proved beneficial for aspects such as detection and classification of malware followed up by DL frameworks deployed for traffic characterization and traffic detection [17-18].

Nonetheless, AI deployment in cybersecurity has several limitations. The major limitations are the difficulties in obtaining data on cybersecurity-related incidents [19], the vulnerability of the AI models to adversarial attacks [20], and the ethical and privacy issues [21].
The main problem with AI models is their "black-box" nature, which complicates the explanation of the reasoning that led to the decisions made by these systems [22]. This opacity can be a major trust and accountability issue because it can be difficult for people to make sense of the cybersecurity decisions made by an AI system. Thus, this means that AI-driven security systems will be prime targets of attacks, rendering them more vulnerable to breaches and cyber threats [23-24].
XAI has come as a Trump card to combat the black-box problem to overcome these challenges pertaining to AI in Cybersecurity. Providing clear, understandable justifications for the decisions taken by AI systems, XAI improves transparency. This allows both users and experts to grasp the logic behind AI-driven outcomes and the key data supporting them, improving the interpretability and trustworthiness of AI-based models in cybersecurity applications [25].

C. An overview of Generative Artificial Intelligence (Gen-AI)

Given the enormous influence that Gen-AI has on many important domains, it is only reasonable to wonder what makes Gen-AI so extraordinary. Gen-AI derives its capabilities from how it processes vast datasets and integrates them into its algorithms. The randomness in output selection, combined with extensive training data, often results in outputs that exhibit creative and human-like characteristics [26]. To find patterns in big datasets, Gen-AI models make use of cutting-edge deep learning methods like Transformers, Variational Autoencoders, and Generative Artificial Networks [27]. These models can use learnt distributions to produce new material after training. The capabilities of Gen-AI are demonstrated by tools like ChatGPT [28] and DALLE [29], which have attracted a lot of attention. OpenAI's ChatGPT is a well-known chatbot that produces a variety of content, such as essays and code, whereas DALLE uses text descriptions to produce lifelike visuals. Although these Gen-AI tools have the potential to completely transform a number of professions, it is yet unclear what the entire impact and hazards will be. Applications of Gen-AI in cyber security include password protection [30-31], Gen-AI text detection in attack, generate adversarial attack examples, Malware and intrusion detection, Simulated attacks, Creating honeypots, security code generation and transfer, and customized Large Language Models (LLM) for security.

The objectives of this paper are as follows:

- Examine state-of-the-art ML and DL approaches for cyber-attack predictions in various types of cyber security environments.
- Provide an in-depth analysis of benchmark datasets, detailing its attributes and suitability of these datasets for various cybersecurity tasks in cyber-attack prediction.
- Analyse the challenges faced by traditional AI models in cybersecurity, especially in terms of interpretability and adaptability to new threats and explore how XAI approaches address these challenges. Provide the insights of how GenAI is used as customized LLM for real-time cybersecurity applications
- A comprehensive examination of the current literature highlights areas for further research and encourages future exploration in the field of cyber-attack prediction.

II. FUNDAMENTAL CONCEPTS OF CYBER SECURITY, CYBER-ATTACKS, AND CYBER-SPACE THREATS

In today's era, safeguarding data through cybersecurity measures is crucial due to escalating cyber risks such as data breaches, ransomware attacks, and identity theft incidents. It is essential for all organizations, irrespective of their size, to prioritize cybersecurity to thwart access or tampering with information. The rapid development of new technologies side by side with the rise of cyber threats create a dilemma for organizations and persons.

Cyber-attacks are the cause for problems like Privacy breaches, Monetary frauds and stealing of the Government property. Hence, it is imperative to understand how cybercrime detection and prevention works. In order to accurately respond to these threats, organizations need mechanisms for exchanging details about attacks and security during an incident response. This helps to resolve security breaches and support the recovery process.

In addition, as devices become increasingly data-driven, organizations need cybersecurity tools that can identify risks before they materialize.

A. Cyber attacks

Cyber attacks are evil attempts to illegally enter computer systems, networks, or data with the intention of stealing, causing physical harm or sabotage. Therefore, learning about the various types of cyber-attacks and how they function is essential for people and businesses to boost their security. These activities are also known as malicious operations performed by individuals or groups that infiltrate computer systems or networks to remove, modify, or erase information, halt services, and achieve other forms of destruction. They can be targeted at different things, including money-making and military or political reasons.

B. Types of Cyber Attacks

i. Untargeted Attacks: In this type of attack, attackers do not have a specific target on the device, service, or user they are attacking. Phishing, waterholing, ransomware, and scanning are some of the techniques used in these types of attacks.

ii. Targeted Attacks: Targeted attacks are explicitly aimed at specific organizations because of their particular interest in financial gain. These types of attacks can be more severe because they exploit vulnerabilities in target personnel or processes. For example, spear-phishing botnets are deployed for DDOS and supply chain subversion.

iii. Insider Threats: This involves employees who launch malicious insider threat activities to breach security systems and steal sensitive information.

iv. Cyberwarfare: For economic or social reasons, governments commit cybercrimes against other countries, resulting in cyberwarfare.

C. Common Cyber Attacks

Many researchers have presented a taxonomy of cyber-attacks with respect to specific attacks and domains [32-35]. Figure 2 depicts the different types of common cyber-attacks.

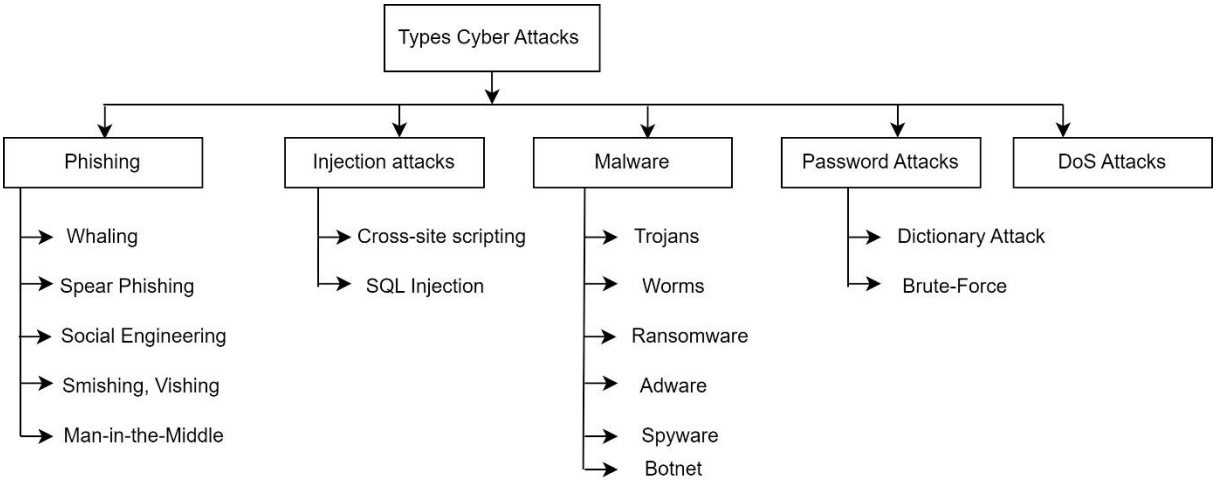

Figure 2. Common types of cyber-attacks

i. Phishing: Phishing attacks where socially engineered e-mails deceive recipients into installing malware or revealing confidential information. Phishing incidents typically aim to obtain access to private and sensitive data such as usernames, passwords, credit card details, and network access credentials. The underlying objective is to persuade the recipient that the message contains valuable or necessary information. Phishing schemes may utilize email, telephone calls, text messages, and social media platforms to deceive individuals into sharing sensitive information [36]. Figure 3 depicts the deceptive and technical subterfuge types of phishing.

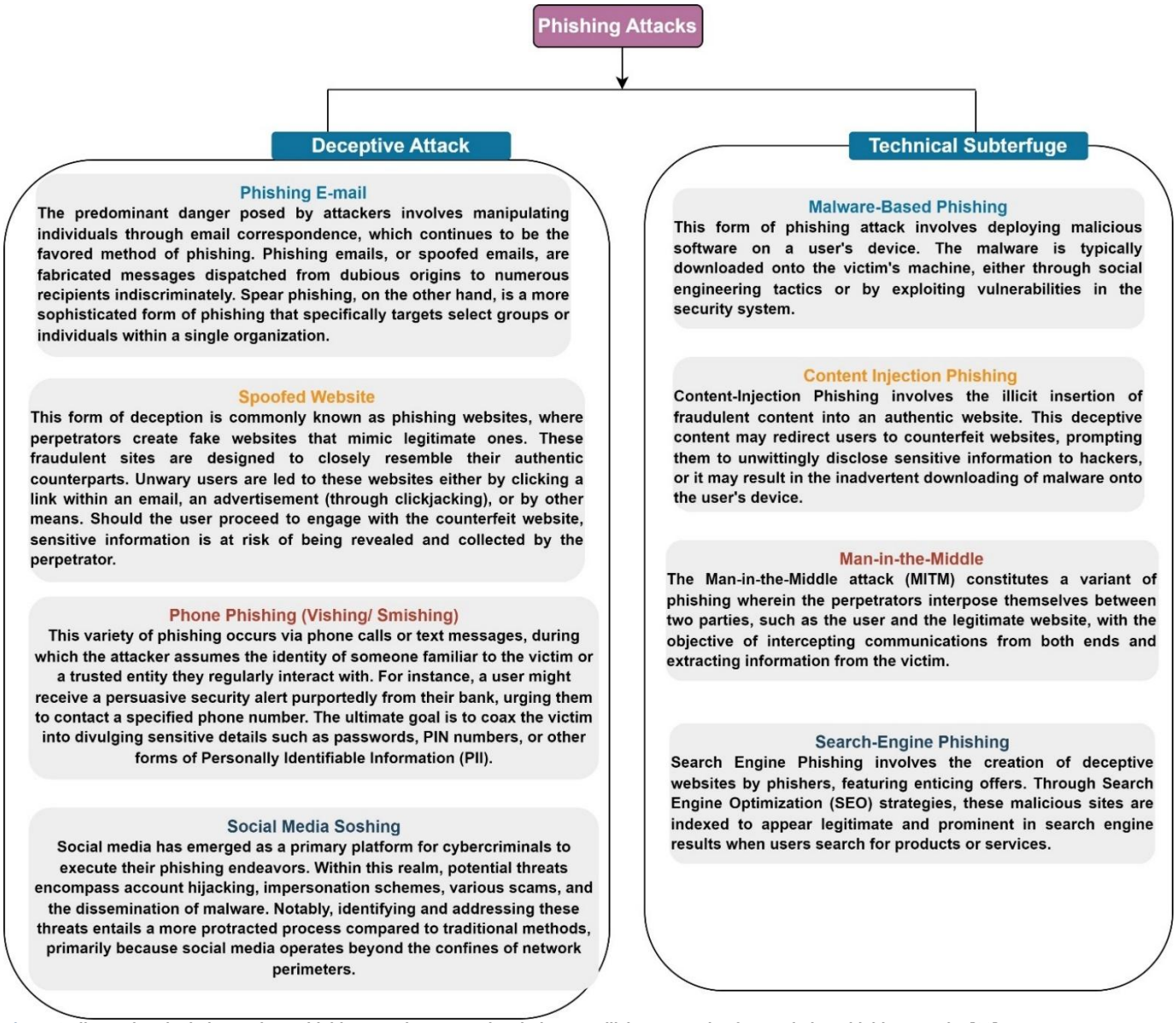

Figure 3. **Illustration depicting various phishing attack types and techniques, utilizing strategies from existing phishing attacks [36].**

ii. Malware-based Attack: This term encompasses malicious software such as ransomware, spyware, and trojans, which may result in data theft, unauthorized access to systems, operational disturbances, or systems becoming non – functional. Malware poses a significant risk to cyber-security. Malware variants are not unique, with a single variant potentially evolving multiple new features. This characteristic makes malware one of the most severe digital threats to cybersecurity [37]. Figure 4 shows the types of malware attacks.

iii. DDOS Attacks: Attacks of Distributed Denial of Service where traffic blocks the system, leading to service interruptions.
iv. Zero-Day Attacks: This occurs when hackers exploit unknown vulnerabilities in software or systems before the manufacturer has an opportunity to correct them.
v. Logic bombs: Malicious code written to perform destructive actions when certain conditions are satisfied.
vi. Abuse tools: Software applications for taking advantage of system weaknesses.
vii. Sniffers: Programs that monitor and capture data passing through a network, such as passwords and other confidential information.

D. Types of Cyber Threat Actors

i. Hostile Nation-States: Because of their advanced capabilities, nation-states pose complex dangers through their cyberwarfare programs, which range from propaganda to the disruption of vital infrastructure.
ii. Terrorist Groups: As they grow more technologically proficient, terrorist groups pose serious concerns as they employ cyberattacks to harm national interests.
iii. Corporate Spies and Organized Crime Organizations: These groups conduct secret trade theft, industrial espionage, company disruption, and cyberattacks with the intention of making money.
iv. Hacktivists: Rather than destroying infrastructure, hacktivists use internet power to further political causes.
v. Disgruntled Insiders: By disclosing private information or infecting systems with malware, insiders, including staff members and outside vendors, pose a frequent threat to cybercrime.

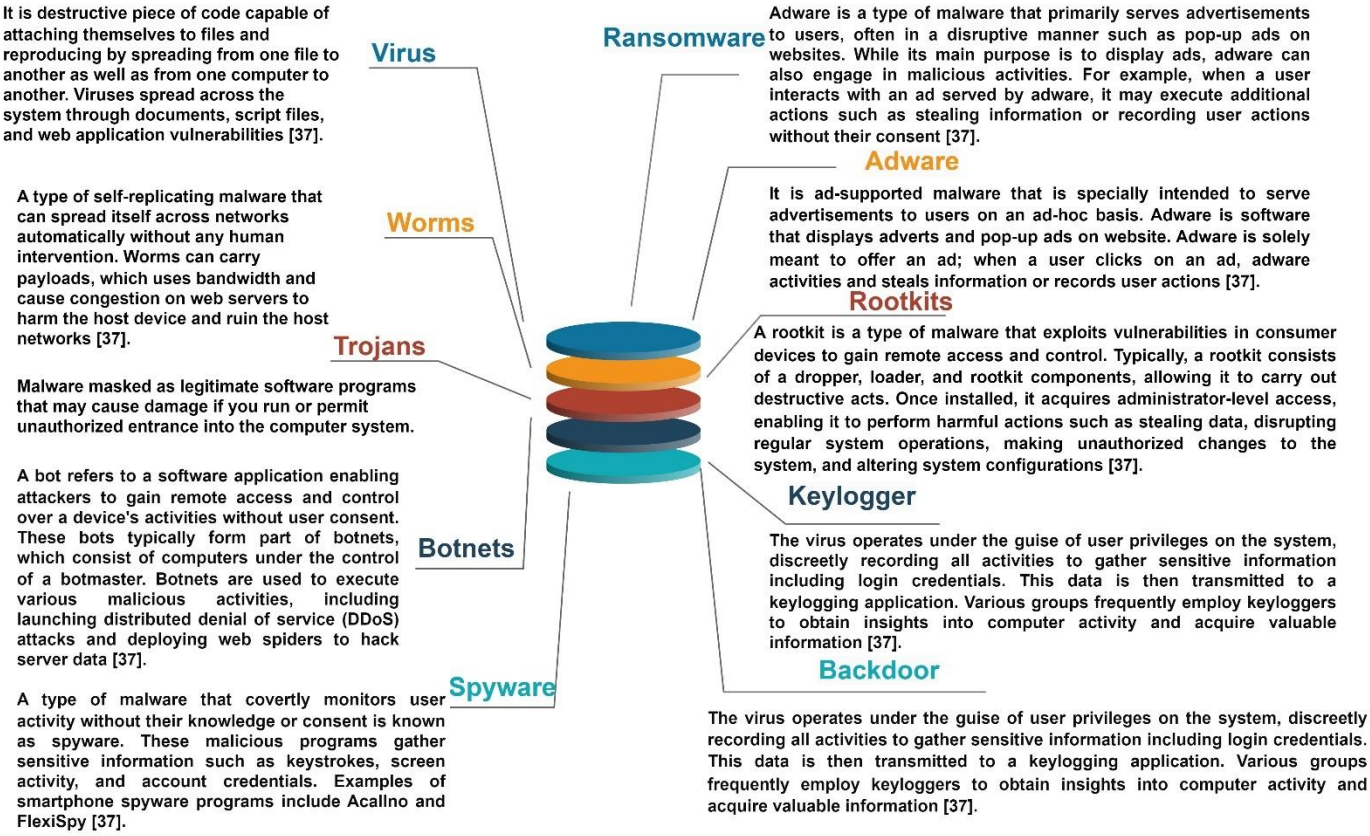

Figure 4. **Types of Malware attacks [37]**

E. Impact of Cyber Threats

Privacy Concerns: Cybercriminals prey on individuals' personal information, resulting in breaches of privacy and monetary damage.

Financial Security: Through various cyber fraud methods, threat actors can steal money, conduct fraud, and disrupt financial systems by obtaining login credentials and personal information.

Economic Health: Cyber threat actions force businesses to incur unwelcome expenditures, such as ransom payments, business interruptions, reputational harm, intellectual property theft, and clientele loss. Operators may reduce risks and safeguard vital services from ever-changing cyber threats by emphasizing cybersecurity resilience, making significant defense investments, and improving threat intelligence sharing.

1. Cryptojacking: Cybercriminals hijack devices to mine cryptocurrency, causing performance issues and downtime for affected businesses.
2. Cyber-Physical Attacks: A major threat to national security is the hacking of vital infrastructure, such as transportation and electricity grids.
3. State-Sponsored Attacks: Nation-states use cyberattacks to breach vital infrastructure and governments, endangering people and private businesses.

III. STATE-OF-ART BENCHMARK DATASETS FOR CYBER ATTACK PREDICTION

Datasets play a significant role in detecting cyber-attacks using ML and DL approaches. There are many datasets that are openly available to researchers for predicting various attacks. Datasets are available for specific attacks and application areas. With respect to this, datasets are classified into seven categories. Figure 5 depicts the types of cyberattack datasets based on the specific application areas [38].

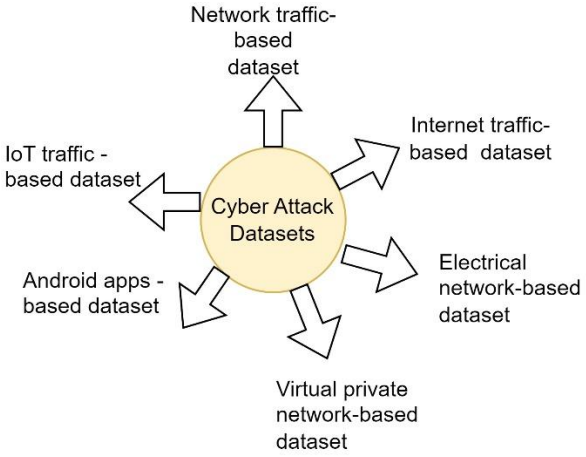

Figure 5. Types of Datasets based on applications.

A. Network Traffic based / Network intrusion detection Datasets.

Publicly available intrusion detection datasets are crucial for effectively comparing the different intrusion detection methods. Additionally, third parties can verify the quality of these datasets only through public availability. KDL Cup 1999 and NSL-KDD are the most used network intrusion detection datasets. Table 1 lists the state-of-the-art benchmark datasets for network-based attacks. From Table 1, SSENET-2014 is the only dataset that is balanced, and the remaining datasets are unbalanced.

Table I.
State-of-the-art benchmark datasets for network-based attacks [39]

Dataset	Year of traffic creation	Data Volume	Duration	Type of traffic	Supervised data?	No. of attacks within the dataset	Attacks considered within the dataset	Limitations
Darpa[40]	1998	NS	7 weeks	Emulated	T	7	DoS, RootKit, Password attack, remote FTP, Nmap, buffer overflow and synflood	Outdated traffic patterns, synthetic data, lacks modern attack diversity
KDD CUP 99 [41]	1998	5 million samples	7 weeks	E	T	4	Four groups of simulated attacks namely DoS, U2R, R2L and research attacks.	Contains redundant data and unrealistic scenarios, outdated attacks
Twente [42]	2008	14M flows	6 days	Real	T	1	Attacks against a honey pot	Limited attack diversity, short duration
CDX [43]	2009	14 GB Packets	4 days	Real	F	NS	-	Insufficient attack variety and data
UNIBS [44]	2009	79k flows	3 days	Real	F	NS	-	Small dataset size, limited duration
ISOT [45]	2010	11GB packets	NS	Emulated	T	1	Botnet	Lack of diversity, emulated traffic

SSENET-2014 [46]	2011	200L points	4 hours	Emulated	T	4	Flooding, port scans, botnet and privilege escalation	Short duration, synthetic traffic
CIC DoS [47]	2012/2017	4.6 GB Packets	24 hours	E	T	1	DoS attacks	Limited attack variety
ISCX 2012 [48]	2012	2M flows	7 days	E	T	4	Dos, DDoS, Brute force and infiltrating the network from the inside	Emulated traffic, lacks modern attack diversity
TUIDS [49]	2012	250 k flows	21 days	E	T	5	DOS, DDoS, port scan, botnet, brute force	Outdated attack types
Booters [50]	2013	250 GB packets	2 days	R	F	1	9 Different DoS attacks	-
CTU-13 [51]	2013	81M flows	125 hours	R	T	1	Botnet attacks	Limited attack coverage
SSHCure [52]	2013/ 14	2.4 GB flows	2 months	R	F	1	SSH attacks	Focused on a single attack type
Botnet [53]	2014	14 GB packets	NS	E	T	2	Botnets and application layer DoS attack	Limited attack diversity
SANTA [54]	2014	NS	NS	R	T	4	DoS, port scan, DNS amplification, and heartbleed	-
AWID [55]	2015	37 M packets	1 hour	E	T	-	Attacks on 802.11	Short duration, limited diversity
IRSC [56]	2015	NS	NS	R	T	NS	-	Insufficient attack information
UNSW-NB15 [57]	2015	2M points	31 hours	E	T	10	DoS, spam, port scan, generic, shellcode, worms, fuzzers, reconnaissance, backdoors, and exploits	outdated attack types
DDoS 2016 [58]	2016	2.1 M packets	NS	S	T	1	Various DDoS attack	Limited diversity
NDSec-1[59]	2016	3.5 M packets	NS	E	T	7	Injection attack, botnet, probe, DDoS, brute force, exploits, spoofing and SSL proxy	Synthetic data, lacks variability
CICIDS 2017 [60]	2017	3.1M Flows	5 days	E	T	8	SQL injection, DoS, DDoS, botnet, cross-site scripting, brute force, infiltration, and heartbleed	emulated traffic
CIDDS-001 [61]	2017	32M flows	28 days	E	T	3	DoS, brute force and port scan	Synthetic traffic, insufficient attack diversity
CIDDS-002 [62]	2017	15 M flows	14 days	E	T	1	Port scan attacks	Limited diversity
Unified Host and Network [63]	2017	150GB flows	90 days	R	F	NS	-	-
UGR'16 [64]	2016	16900M flows	4 months	R	T	5	DoS, Spam, port scans, brute force and botnet	Large dataset, computational complexity
LITNET 2020 [65]	2020	1.35 GB	NS	R	T	12	Scan, spam, reaper worm, fragmentation, code red, land, flood (HTTP, UDP, ICMP, SYN), smurf, W32. Blaster	Recent attacks, but small dataset

HIKARI-2021 [66]	2021	NS	39 hours	R	T	4	Bruteforce, bruteforce-XML, probing and XMRIGCC CryptoMiner	-
ROSIDS23 [67]	2023	NS	NS	Realistic	T	4	Subscribing Flood, DoS, Unauthorized publish and subscribe	-
RoEduNET2021 [68]	2021	6,570,058 frames of pure traffic and 5,637,815 flows that are labeled as anomalies.	NS	NS	T	2	DoS and portscan attack	-
MSCAD [69]	2022	NS	NS	R	T	2	Password attacks, volume-based DDoS, App-based DDoS, Web Crawling, and port scan traffic	-

B. Malware and Android app based datasets.

Malware datasets play a major role in cybersecurity research. Many cyber-security researchers have generated benchmark malware datasets to study the vulnerabilities exploited by various malware, benchmark the effectiveness of security tools, and provide information about emerging threats and malware families. Table II lists some of the malware datasets.

Table II.
State-of-the-art benchmark datasets for Malware and Android app based datasets.

Dataset	**Malware Types included in Dataset and samples**																	**Limitations**
	Spyware	**Trojan**	**Worms**	**Adware**	**Dropper**	**Virus**	**Backdoor**	**Downloader**	**Dataeraser**	**Ransomware**	**Zero-Day**	**Riskware**	**Scareware**	**smsware**	**Generalmaware**	**Flooder**	**DoS/DDos**	
Mal-API-2019 [70]	832	1001	1001	379	891	1001	1001	1001	-	-								Limited malware types, no modern sophisticated threats
CCCS-CIC-AndMal-2020 [71]		13559		47201			1538			6202	13,340	97,349	1,556					-
CIC-AndMal2017 [72]	-	-	-	104	-	-	-	-	-	101	-	-	112	109				-
CIC-AAGM2017 [73]				250 apps											150 apps			Limited sample size
SoReL-20M [74]	4550007		3414132	2411262	3577111			2565838		1152354						101595		-
Alibaba Cloud Malware Detect			100			4279	515			502							820	Insufficient diversity, focused on limited attack types

ion [75]																		

C. IOT-Traffic-based datasets.

The Internet of Things (IoT) is gaining popularity and rapid development, which has led to a wide range of issues for both manufacturers and users. One of the main concerns is the security of IoT applications and devices [76]. There are various ways to acquire network datasets relevant to the IoT. In a testbed-driven generation, researchers instrument an IoT device environment and record the network traffic in normal and attack scenarios to create datasets. This is a labor-intensive task that requires resources in the form of money (for technology) and time (for data collection). However, because synthetic dataset generation relies on the modeling or emulation of IoT devices, communication networks, and apps that operate on top of them, they use fewer resources. Although this strategy is incredibly versatile, it can be challenging to obtain components that behave realistically. Ultimately, network traffic from actual IoT devices used by consumers is recorded to provide empirical datasets [77]. Numerous researchers have compiled datasets aimed at detecting attacks on IoT traffic. The process of generating these datasets had several characteristics. First, IoT traffic can be categorized based on its type, such as whether it is IP-based or specific to IoT. Additionally, the traffic data content can vary, including full packets, headers, features, sensor data, or signal data. Furthermore, the scale of the dataset is a crucial aspect that encompasses the number of devices involved and the duration of data records. The dataset may also be categorized based on the use of the devices, which can range from smart home applications to health monitoring, wearables, and Wireless Sensor Networks (WSN). Finally, the methodology used for data collection contributes to the dataset characteristics, which may include real-world data collection, simulation, testbed experiments, emulation, or a hybrid approach. Table III presents state-of-the-art IoT traffic-based datasets.

Table III.
State-of-the-art IoT traffic-based dataset

Dataset	Year of traffic creation	No. of attacks within the dataset	Attacks considered within the dataset	Features	Limitations
IOT-23 Dataset [78]	2020	9	FileDownload , DDoS, Okiru, Torii, CC, Part Of A Horizontal, Mirai, PortScan, HeartBeat,	23 Packet related features and 44 addresses and port related features	Limited attack diversity, lacks advanced IoT-specific threats
TON-IoT [79]	2020	9	DoS, DDoS, password attack, Ransomware, scanning, data injection, backdoor, Cross-site Scripting (XSS), and man-in-the-middle.	44	Focus on simulated attacks, may not represent real-world IoT attack patterns
MQTT-IoT-IDS2020 [80]	2020	2	Aggressive scan (Scan A), UDP scan	44	Limited attack types, lack of complex threat scenarios
Edge-IIoT [81]	2022	5	DoS and DDoS attacks, information gathering, man-in-the-middle attacks, injection attacks, and malware-based attacks	1176	-
N-baiot [82]	2018	10	Malicious attacks from two botnets	24	
WUSTL-IIOT-2021 [83]	2021	4	Backdoor, DoS, command injection and reconnaissance	41	
X-IIOTID [84]	2021	18	Reconnaissance, Ransom DoS, weaponization, exploitation, Command and Control, lateral Movement, tampering, crypto-Ransomware	59	
Bot-IoT [85]	2019	6	DoS, DDoS, Service scanning, OS Fingerprinting, Key logging and data theft	46	

D. Virtual private network-based dataset

This dataset, ISCXVPN2016, is proposed by authors in [86] and it is composed of standard and VPN-based network traffic. The dataset is labeled and consists of diverse network activities including web browsing (Firefox), e-mail (SMPTS), chat (Skype), streaming (YouTube), file transfer

(SFTP), VoIP (Hangouts voice calls), and peer-to-peer (uTorrent).

The dataset, referred to as CIC-Darknet2020, was released by the authors of [87] in 2020, and includes features from traffic captured from two darknets, namely The Onion Router (Tor) and a virtual private network (VPN). The dataset includes 158,659 samples with hierarchical labels, where the 1st layer labels for traffic category are Tor, non-Tor, VPN and non-VPN.

E. Electrical network-based datasets

LBNL [88], IEEE 300-bus power test system [89] and ICS cyber-attack datasets [90] are electrical network-based datasets used for cyber security [38]. The LBNL dataset was gathered using the uPMU (micro-phasor measurement unit) at the electrical network of Lawrence Berkeley National Laboratory. The uPMU generates 12 data streams at a frequency of 120 Hz, providing high-precision measurements with timestamps accurate to within 100 nanoseconds. This dataset is applicable for tasks such as microgrid synchronization and the characterization of loads and distributed energy generation [38, 88].

Authors of [89] provided the information about IEEE 300-bus power test system. This dataset provides a topological and electrical structure of power grid, which is used especially for the detection of false data injection attacks in the smart grid. The system has 411 branches, and average degree (< k >) of 2.74.

The ICS datasets comprise of five distinct components: (1) Power System Data, (2) Gas Pipeline Data, (3) Energy Management System Data, (4) New Gas Pipeline Data, and (5) Gas Pipeline and Water Storage Tank Data. The Power System dataset includes 37 scenarios categorized into 8 natural events, 1 no-event scenario, and 28 attack events. The attack events are further classified into three types: (1) relay setting changes, (2) remote tripping command injections, and (3) data injections. These datasets are valuable for cybersecurity intrusion detection within industrial control systems [38, 90].

F. Internet Traffic-based datasets

These datasets focus on broader internet traffic, often from ISPs or cloud platforms, capturing a wide range of activities. UMASS dataset [91], Tor and non-Tor dataset [92] and MAWI Working Group Traffic Archive [93] are examples of internet traffic-based datasets.

UMASS dataset[91] comprises two components: simple timing attacks on OneSwarm and strong flow correlation attacks. The simple timing attack on OneSwarm complies with the constraints of general criminal procedure. It includes three types of attacks: timing-based, query forwarding-based, and TCP throughput-based attacks. The strong flow correlation attacks involve multiple Tor clients browsing the top 50,000 Alexa websites via Tor.

Authors of [92] proposed the Tor-nonTor dataset. This dataset features eight categories of network traffic: VOIP, chat, audio streaming, video streaming, email, P2P, browsing, and file transfer. It includes data collected from over 18 widely used applications, including Spotify, Skype, Facebook, and Gmail.

The MAWI dataset [93] comprises daily traffic traces in the form of packet captures, collected from a trans-Pacific link between Japan and the United States. This dataset is valuable for researching anomaly detection, analyzing internet traffic patterns, and developing traffic classifiers.

There are some limitations in the benchmark datasets discussed in this sections that can potentially impact the model generalizability and reliability of AI:

Narrow attack scope: Datasets, such as MQTT-IoT-IDS2020 and N-baiot, provide few classes of attacks that restrict AI models' generalization on heterogeneous attacks in real-world scenarios. Specific emerging IoT threats are absent in even broader data sets like IOT-23.

Obsolete or Simulated Attack Patterns: Datasets created in older times such as Bot-IoT (2019) and N-baiot (2018) do not cover the latest attacking techniques. Moreover, datasets like TON-IoT are based on simulation traffic that does not cover the complex scenarios of real-world attacks.

Table IV summarizes the classification of various cyber security datasets.

.

Table IV
Classifications and key features of cyber security datasets

Dataset Type	Description	Purpose	Examples	Key Features
Network Traffic-Based Datasets	Capture traffic data within a network, including packet flows, IP addresses, port numbers, and protocols.	Detecting network intrusions, DDoS attacks, and abnormal network behavior.	CICIDS2017, UNSW-NB15, LITNET 2020	Rich in network packet details; useful for traffic pattern analysis
Internet Traffic-Based Datasets	Focus on broader internet traffic, often from ISPs or cloud platforms, capturing a wide range of activity.	Identifying large-scale cyber-attacks, web threats, and suspicious behaviors.	UMASS dataset, Tor and non-Tor dataset and MAWI Working Group Traffic Archive	Captures diverse internet interactions; useful for detecting large-scale internet threats.
Electrical Network-Based Datasets	Collected from electrical and smart grid networks, monitoring data flow in critical systems.	Detecting threats targeting power grids and industrial control systems	Power system datasets,	Focus on stability and control signals; useful for preventing attacks that cause physical disruption.

Virtual Private Network (VPN)-Based Datasets	Focus on traffic routed through VPNs, capturing encrypted communication patterns.	Detecting malicious activities and misuse within encrypted traffic.	VPN-Filter-related datasets	Encrypted data patterns; challenging to analyze, useful for identifying suspicious VPN behavior.
Android Apps-Based Datasets	Contain data from Android applications, such as permissions, system calls, and network requests.	Mobile malware detection and protecting Android devices.	Drebin, Android Malware Genome Project	Focus on app behavior and permissions; essential for malware detection in Android environments.
IoT Traffic-Based Datasets	Data from IoT devices, generating small and frequent packets due to limited processing power.	Identifying threats targeting IoT devices like botnets and spoofing.	UNSW Canberra IoT datasets, Bot-IoT dataset	Device-specific data; lightweight, crucial for detecting attacks on IoT systems
Internet-Connected Devices-Based Datasets	Broader datasets from internet-connected devices, such as PCs, servers, and smart appliances	Identifying malware, unauthorized access, and threats across connected devices	Various enterprise network datasets	Heterogeneous data from diverse devices; suitable for detecting multi-platform attacks

IV. Role of Machine Learning algorithms (ML) for Cyber Attack prediction

The application of ML in cybersecurity shows considerable promise for strengthening security systems and protecting against cyberattacks. To keep up with the ever-changing nature of cyber threats, it is crucial to create and refine techniques continuously [94]. Few likely solutions based on machine learning are vulnerable to adversarial assaults, highlighting the need to consider this weakness when developing countermeasures for sophisticated cyber threats.

A. Importance of ML in cyber security

ML algorithms [95] can process large amounts of structured and unstructured data, extract valuable patterns, learn from past data, and predict outcomes accurately. ML-based systems can help analyze possible hazards and threats within a firm, aiding in risk assessment and cybersecurity planning [96] owing to their learning and pattern-finding capabilities. ML is becoming a prominent tool in cybersecurity. As the number of large-scale cyberattacks increases, cyber security professionals require faster and more accurate threat identification and prevention. Machine learning is an intriguing approach.

B. Types of learning approaches

Several ML-based methods are used in cybersecurity, including regression, probabilistic models, distance-based learning, decision trees, dimension reduction algorithms, and boosting and bagging techniques. These machine-learning technologies help detect data breaches and vulnerabilities in computer systems and networks. One major feature is the ability to evaluate and alter large amounts of data without relying on subject specialists. Machine learning techniques can be broadly divided into supervised, unsupervised, semi-supervised, and reinforcement-learning techniques. Figure 6 depicts the ML workflow for cybersecurity.

i. **Supervised Machine Learning**: This refers to algorithms that require developer supervision. The developer tags the training data and establishes stringent rules and constraints for the algorithm. Algorithms can use labelled examples [97] to predict future events by applying the knowledge gained from previous data to new information. The supervised approach forecasts the target variable by using a function created over several inputs. Audited algorithms identify the input data and the intended results.

ii. **Unsupervised Machine Learning:** Unsupervised ML approaches are used when training data lack labelled data or classification. This learning technique explores how computers extract functions from un-labeled inputs to reveal hidden structures [98]. Unsupervised techniques may detect all types of cyberattacks, including undiscovered ones, by identifying system irregularities. Unsupervised machine learning (ML) is commonly used in cybersecurity to detect anomalies, IoT-based zero-day attacks [99], classify entities, and explore data.

iii. **Semi-supervised Machine Learning:** Using a combination of labelled and un-labeled data can improve learning precision [98]. The semi-supervised technique efficiently detects new cyber-attacks by identifying abnormalities and applying them to other types of attacks. It can be used to identify network breaches, DDoS attacks, and malware.

iv. **Reinforcement Learning:** The algorithm evolves and chooses the best strategy through iterative processes. Machines and software agents can use this process to automatically determine the best behaviors to maximize performance under a particular circumstance [100]. RL is helpful for system penetration testing, risk assessment, and the identification of aberrant behaviors.

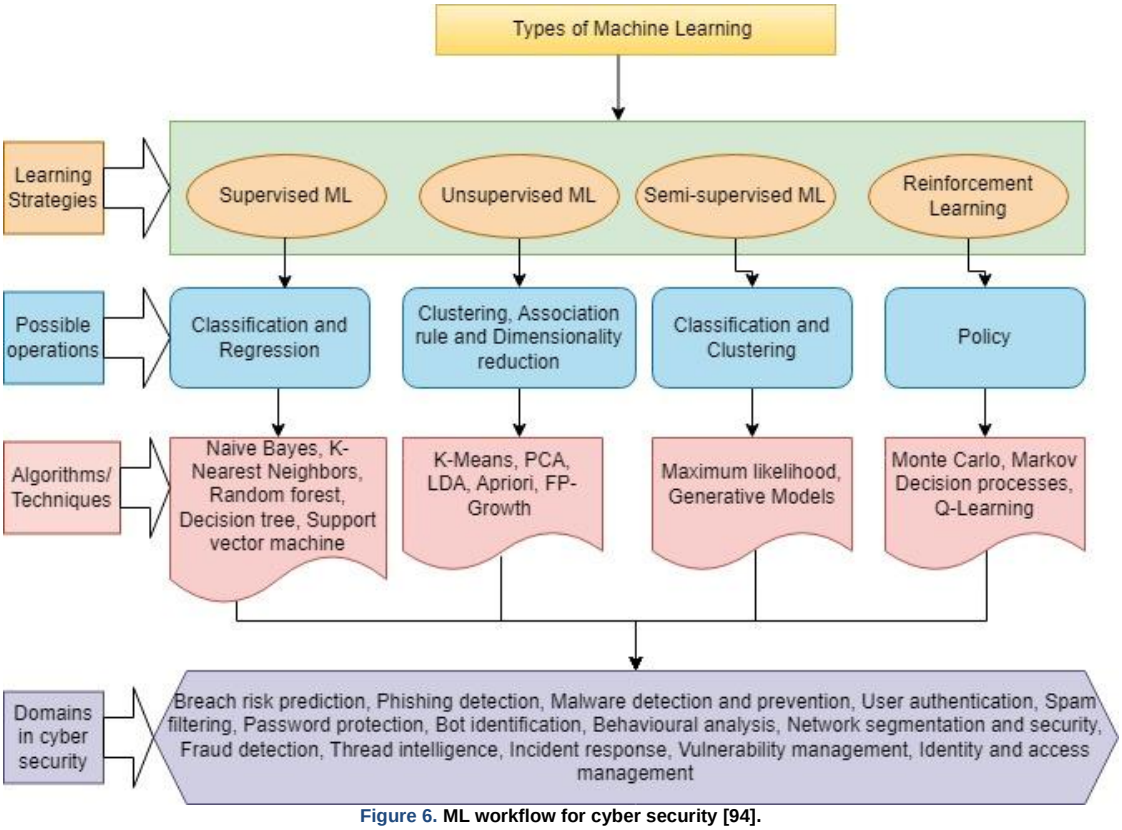

Figure 6. ML workflow for cyber security [94].

C. Types of ML Algorithms

ML techniques have been used in numerous cybersecurity applications. These methods include regression, classification, clustering, dimensionality reduction, and boosting.

Regression analysis predicts continuous values based on the independent variables given, and algorithms to perform this include simple and multiple regression, which require one and multiple independent variables, respectively, to predict dependent variables. Polynomial regression analyses the relationship between dependent and independent variables in a polynomial degree form. LASSO and Ridge regression [101] are popularly known as effective approaches that are typically employed for developing learning models in the presence of a high number of features, as they are capable of preventing overfitting and decreasing the model complexity. Regression classifiers are used to detect fraud, malware, and other types of attack.

Classification techniques predict discrete values (binary/multiple) based on the features fed into a model. The naive Bayes classifier assumes that its features are independent of each other. It works best with a small amount of data but can handle noisy data. Logistic regression works well with linearly separable data points based on the calculated probability. Decision Tree [102] method is a non-parametric method. Here, the most relevant feature becomes the root node, the branch nodes hold the features, and the leaf nodes are the classes. This splitting or construction of the tree can be achieved using the entropy/Gini index criteria. The random forest method uses majority voting or the aggregate method to select the outcome from several decision trees constructed in parallel over subsamples of data. The support Vector machine finds the optimal hyperplane that represents the margin separation between classes.

Clustering Analysis divides data points into clusters that are more similar to one another than to the other groups. This is achieved using unsupervised machine learning techniques. One such popular technique is K-means clustering, which is most suited when data samples are well distributed, based on Euclidean distance clusters formed until there is no change in group assignment. Another important clustering is agglomerative hierarchical, wherein data samples are initially considered to be singleton; later, they are slowly paired up and finally form a single cluster using single/complete/average linkage.

Association rules help build relationships between predictors with statements like 'IF' and 'THEN.' Suppose that a person buying bread in a supermarket is more likely to buy jams along with it. Apriori [103] is the most commonly used technique that requires knowledge of frequent item-set characteristics and generates candidate item sets. It uses a support and confidence approach to determine the rules. FP-growth [104] rules based on a frequent-pattern tree using the divide-and-conquer method.

Dimensionality Reduction involves feature selection and extraction. Feature selection is an important phase in which the most relevant independent variables are chosen from the original dataset, which, in turn, reduces the model's complexity and overfitting. This can be performed using chi-square, ANOVA [105], Pearson's correlation coefficient, and recursive feature elimination techniques. Feature extraction reduces features from the original dataset by extracting new features and ignoring unimportant features. This phase helps to better understand the data. In Principal Component Analysis (PCA), new brand components can be created by extracting low-dimensional space from the current dataset attributes.

Policy-based techniques can be employed through reinforcement learning. In this type of ML, the agent interacts with an unknown environment. Each action receives a reward in terms of positive/negative. Actions extracted with the maximum positive rewards are said to be the optimal policy in the RL environment. When model dynamics, such as transition probability, rewards, and the next state, are given, it is called a model-based approach. A Markov decision process can be used to solve this type of problem. When model dynamics are not given, model-free techniques, namely Monte Carlo, Q-Learning, SARSA, and Deep Q-Learning [106], are used. Real-world applications of RL include game theory, control theory, operations analysis, information theory, simulation-based optimization, manufacturing, supply chain logistics, multi-agent systems, swarm intelligence, aircraft control, and robot motion control [107].

All these ML algorithms are used to predict cyber-attacks. Table V lists the state-of-the-art ML techniques utilized for cyber-attack prediction.

Table V.
State-of-art research in cyber-attack prediction using ML.

Ref. & Publication Year	Purpose	Feature combination and selection	Dataset	Approaches and Performance	Advantages	Scope of enhancement
[108] 2018	Multi-classification of malware in to five categories using corpus of malware analysis reports.	Features obtained in JSON format of CUKKOO sandbox malware report analysis	Malware samples were gathered from online repositories	Obtained average accuracy of 89% with ML techniques such as Decision tree, Gaussian NB, Gradient boosting, KNN, Logistic Regression, Random Forest, SVM	Cross-validation techniques have improved the overall accuracies of ML models used here. This framework can be used to classify the malwares in an organization.	Larger datasets can be used to improve the accuracies and Latest technologies can be employed to improve malware analysis report.
[109] 2023	Malicious attack detection system in real time over encrypted traffic at application level.	Use selection stability and selected predictive techniques for every server.	Datasets are collected from various sources of internet and corporate like CVE-2012- 1652, CVE-2015- 0610, CVE-2008- 2382, CVE-2010- 5304 etc.	Elastic Regression method (ENetRM) obtained precision of 0.946, recall of 0.9764 of recall, F1-score of 0.9221, the throughput of 97.04% and computation overhead of 53.48%.	Compared to standard approaches such as DBSCAN, NAHSMM, NIDS methods, proposed ENetRM obtained increased throughput and reduced computational overhead.	The next work focuses on expanding the architecture to include protocols such as securing remote management connections between cloud-based virtual machines.
[110] 2021	ML based solution for the protection of IoT device.	DDoS taxonomy was used to select the features	Distributed Denial of Service attacks Evaluation Dataset (CICDDoS2019)	Logistic regression obtained an accuracy of 99.7%	Classification of malicious attacks from that of normal namely Benign, NetBIOS LDAP attacks.	Further various datasets and ML techniques can be used for the protection of IoT from cyber-attacks.
[111] 2023	Various ML prediction models, Traffic attack detection is eased, and gather data about the botnet, and send it to a firewall to block suspicious traffic and prevent the attack.	Feature importance score and Extra tree classifiers were used to find the most relevant features.	Real traffic data obtained from Kaggle.	KNN, Random Forest and Naïve bayes classifiers were used, among them Random Forests yielded highest accuracy and precision of 0.97 with less processing time.	Prediction models identified all attacking botnets as well as helped in simulating DDoS attacks.	Scope to explore other ML classifiers.

[112] 2021	An efficient approach to detect advanced persistent threats.	Feature extraction: MCA, PCA and MFA. Feature selection: IG, GR and OneR.	APT dataset	Naïve bayes achieved 91% of highest accuracy among Bayes net, KNN, RF and SVM.	Effective way of feature extraction and selection was used to match APT with Cyber kill chain cycle.	Much larger datasets can be incorporated to detect more cyber threats.
[113] 2023	Brief analysis to know about the keylogger attacks and their detection using Machine Learning.	Recursive feature elimination is used to extract characteristics of key.	Keylogger data	SVM in ML has proved its efficiency for the detection of keylogger attacks.	Machine learning based detection outperformed than Firewall based detection.	Further there is a scope to provide robust solution for various types of keylogger attacks and also for detecting them.
[114] 2021	This study compares ML algorithms for detecting attacks and anomalies based on a detailed review of Machine Learning approaches and the importance of IoT security against various forms of attacks.	Various unsupervised and reinforcement learning techniques were used for feature reconstruction.	UCI, IoT-23, BoT-IoT, NSL-KDD	Combining RF and DT machine learning algorithms improved attack detection accuracy. In studies, two ML algorithms, RF and KNN, reached 99% accuracy in detecting attacks.	This study focuses on integrated Machine Learning algorithms for IoT security, providing a comprehensive review of IoT attacks and their implications.	Research on Machine Learning algorithms has identified potential issues that can assist future researchers in achieving their aims in this subject.
[115] 2021	Analysis of IoT attacks and detection using ML techniques	pcap libraries, Random forest and Bagged trees were used in feature extraction.	Live Dataset is generated by ThingSpeak IoT Cloud.	SVM, RF and Bagged Trees (BT) were used for IoT attacks wherein BT performed well for huge datasets.	Dataset was generated from live network; this helps in finding attacks of different domains.	Further different ML classifiers with various hyper-parameters can be used to improve the efficiency of prediction models.
[116] 2022	This research developed a new method for detecting and identifying cyber-attacks on electric grid load prediction data.	K-means, mean-shift and hierarchical algorithms are used to obtain scaling data	Aguilar Madrid, Ernesto (2021), "Short-term electricity load forecasting (Panama case study)", Mendeley Data, V1, doi: 10.17632/byx7sztj59.1	To perform classification, DT, RF, Gaussian NB, Gradient boosting, SVM, KNN were used, RF outperformed among these with 95% accuracy.	To address the classification of different types of attacks, a hybrid model approach was introduced.	Future research could improve by analyzing the sensitivity of various load forecasting systems to cyber-attacked data.
[117] 2023	This paper uses deep learning techniques to classify cyber-attacks and a metaheuristic algorithm is used to optimize data features.	A restricted Boltzmann machine was used for feature learning and dimensionality reduction.	Data collected from the RTU and other IED components	CNN, ANN, SVM, RBM-RF techniques were used for binary, three, multi-class classification	The results show that the suggested RF-RBM method is effective for detecting and classifying cyberattacks in SCADA systems for smart grids.	Further, CNN architecture can be explored.
[118] 2022	This study analyzed 352 real-life cyberattacks on healthcare organizations using CVSS data to identify trends and specific attacks.	A subset of features selected for the evaluation	CVSS v3 base score estimates were derived from the Cybersecurity and Infrastructure Security Agency (CISA). The dataset comprised observations from January 1999 to May 2022.	Various techniques such as LR, KNN, ANN, DT, RF and linear regression were used to estimate success rates. KNN obtained the highest rate of 87%.	Based on the results, Cyber-attacks pose a significant risk to healthcare establishments.	Further, larger datasets of various nations can be explored.

[119] 2022	This research proposes offloading ML model selection to the cloud and real-time prediction to fog nodes.	Cloud and Fog layer separation of data.	NSL-KDD dataset	For the evaluation, various combinations of base classifiers were used, namely RF, DT, KNN, LR, NB etc.,	KNN, NB and DT combination produced the highest kappa, F-measure and ROC.	Analysis can be further continued with real testbed emulation.
[120] 2019	ML based solution to identify the cyber-attacks in IoT networks.	CICFlowMeter was used to extract the features. Random forest regression technique was used to select the features.	BoT-IoT dataset	NB, QDA, RF, ID3, Adaboost, MLP and KNN were used to evaluate real time traffic data.	NB yielded less processing time and KNN obtained accuracy of 99%.	Further, other machine learning algorithms can be used to improve the performance.
[121] 2023	Detection and classification of cyber-attacks through ML models.	Crossover and mutation in genetic algorithm	IoT traffic real data	Classification and detection of cyber-attack in cyber physical system.	ML techniques were used to employ an intrusion detection system.	Further, other ML algorithms can be used to improve the performance.
[122] 2022	ML solution to enable cyber security systems to detect and prevent terrorist acts while reacting to changing behavior.	Standardization and various plots to understand the correlation among features.	Cyber-attack data	Statistical regression, Random DT, DT and KNN were used to perform cyber-attack classification. KNN outperformed with 98%.	The research aims to help authorities prevent human trafficking and improve cybercrime detection.	Further, other MLalgorithms can be used to improve the performance.
[80] 2020	Generation of MQTT-IoT dataset and evaluation of ML algorithms	Packet, unidirectional and bidirectional flow-based features	MQTT-IoT-IDS2020	Classification accuracy using Bidirectional features. LR- 99.44% k-NN – 99.9% DT- 99.95% RF- 99.97% SVM (RBF Kernel)- 96.61% NB- 97.55% SVM (Linear Kernel)– 98.5%	This study investigated the various obstacles and prerequisites involved in constructing IDS tailored for IoT networks, with a specific focus on an MQTT network as a prime example. MQTT dataset was evaluated using 6 ML algorithms	Results can be interpreted using XAI approaches.
[84] 2022	Creation of the X-IIoTID intrusion data set, a connectivity- and device-independent intrusion data set that fits the heterogeneity and interoperability of IIoT systems.	connectivity-agnostic features	X-IIOTID	DT- 99.49%	This study investigated 5 ML algorithms on novel X-IIoTID dataset and demonstrated that DT outperforms other state-of-the-art ML algorithms for predicting 9 attacks of X-IIoTID dataset.	Results can be interpreted using XAI approaches.
[83] 2019	Proposed ML-based Intrusion Detection system	23 network flow features	WUSTL-IIOT-2021	LR- 99.90% KNN-99.98% SVM- 99.64% NB- 97.64% RF- 99.9% DT- 99.98% ANN- 99.64%	Evaluated 6 ML algorithms for intrusion detection.	Interpretation of results of ML algorithms using XAI approaches
[85] 2019	Design a new realistic Bot-IoT dataset and evaluation using ML and DL approaches	Filter based feature selection	BOT-IOT	SVM- 99.98%	Generation of new dataset and achieved good performance using ML and DL approaches	Optimization of ML and DL methods

D. ML key challenges in cyber security

ML has immense potential for improving cybersecurity defenses; however, it faces numerous significant hurdles in detecting and mitigating attacks. A few of these are as follows:

Data Quality and Quantity: To properly train models, machine learning algorithms require large amounts of high-quality [108] data. The scarcity of labelled cybersecurity datasets makes it challenging to collect labelled data for training purposes. Furthermore, data quality issues, such as imbalanced datasets (in which particular types of data are underrepresented), might impair model accuracy.

Adversarial Attacks: Adversarial attacks are designed to trick ML models by exploiting the flaws in the underlying algorithms. Adversaries in cybersecurity may use tactics such as adversarial examples, which involve modest, carefully engineered modifications to the input data that cause ML algorithms to misclassify them [112]. Adversarial assaults pose a substantial threat to the dependability and robustness of ML-powered cyber-security systems.

Concept Drift: As cyber threats evolve, the underlying data distribution shifts over time. This tendency, known as idea drift, can undermine the performance of ML models trained on historical data, making them less efficient in recognizing new and emerging risks. Adapting ML models to deal with concept drift while preserving their effectiveness over time is a critical challenge in cybersecurity [109].

Interpretability and Explainability: ML models employed in cybersecurity frequently lack interpretability and explainability, making it difficult for security analysts to understand the logic behind model predictions. Interpretability [111] is critical for trust and accountability, as analysts must comprehend why the ML model makes a specific decision to take necessary action. Ensuring that ML-based cybersecurity systems are transparent and understandable is a difficult task.

Resource constraints: Many machine learning techniques, particularly deep learning models, require significant computer resources for training and inference. Deploying and executing complicated ML models in resource-constrained environments, such as edge devices or IoT devices, may be impossible because of processor power, memory, and energy usage limits. Developing lightweight and efficient machine learning algorithms that can be deployed in resource-constrained contexts is a cybersecurity challenge [113].

Privacy Concerns: ML models trained on sensitive cybersecurity data may unintentionally divulge sensitive information or harm user privacy. Federated learning and differential privacy techniques seek to overcome these challenges by allowing collaborative model training across remote data sources, while maintaining privacy. However, ensuring strong privacy protection while preserving the model performance remains a challenge in ML-based cybersecurity systems [118].

V. Role of DL in Cyber Security Domain

Researchers have proposed solutions using deep learning algorithms to detect threats, anomalies, malware and network intrusions, phishing or spam attacks, website defacements, vulnerability assessments, analyzing cyber threat intelligence, user behavior, etc.

a. Importance of DL in cyber security

DL is an area of machine learning [123] that uses multilayer transformations to analyze large volumes of data, find complicated patterns, and generate accurate predictions. In cybersecurity, deep-learning-based defense systems automate cyber-attack detection and continuously improve their capabilities. This allows firms to detect, respond to, and mitigate cyber-attacks more effectively. Its ability to respond to emerging threats and automate security operations makes it a must-have tool in the current cybersecurity world.

b. Various DL models

DL models are broadly classified as supervised and can be applied when labelled data are given for classification and regression tasks. Unsupervised methods are mostly used for representation learning, and self-learning techniques help in feature extraction. As per the learning strategies, deep learning models were mentioned, as shown in the figure 7.

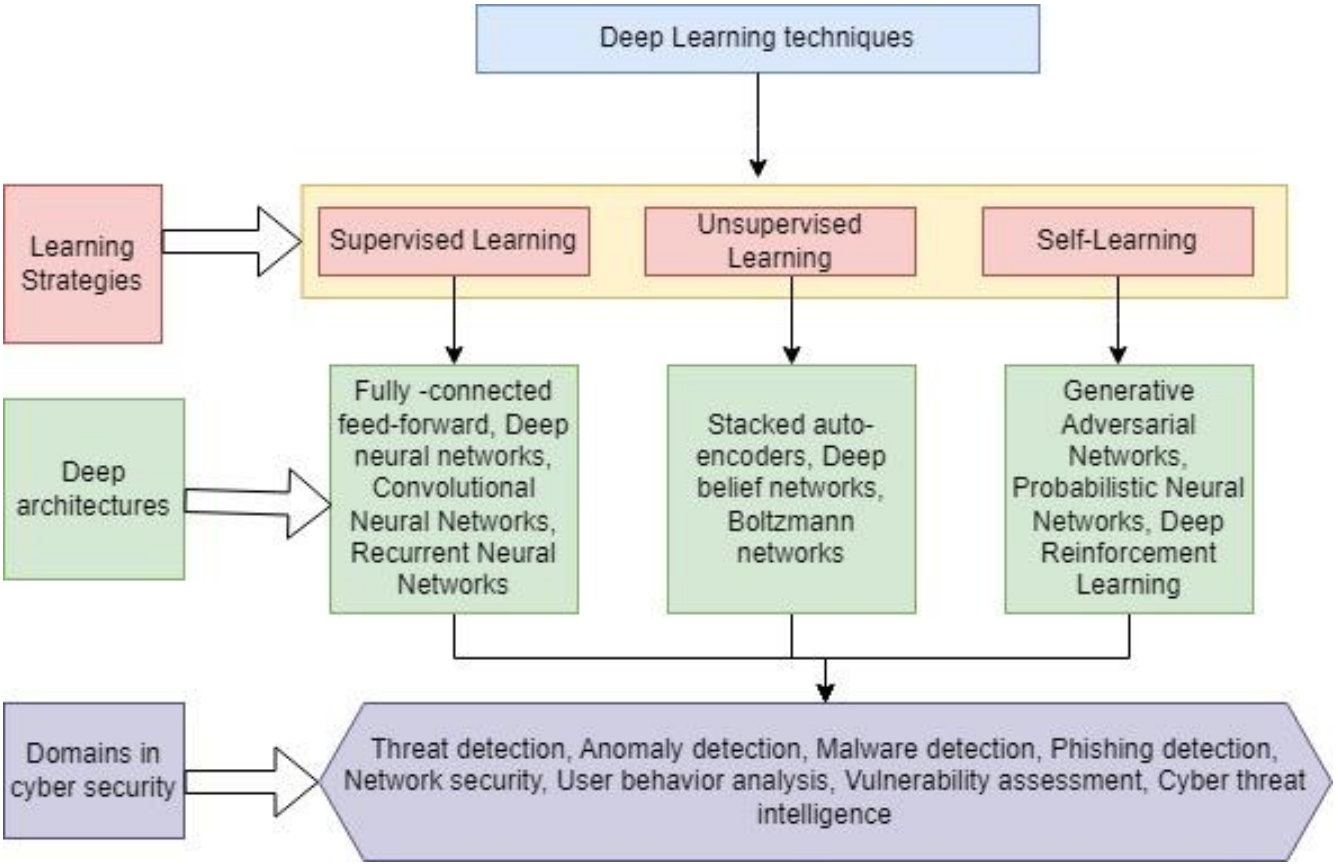

Figure 7. Deep Learning Architectures for cyber security

The prediction in DL models is based on Artificial Neural Networks. An artificial neural network (ANN) is a structure of interconnected neurons that transfers information to one another. DNNs [124] differ from single-hidden-layer neural networks owing to the larger number of hidden layers involved in pattern recognition. A deep neural network (DNN) comprises an input layer, many hidden layers, and an output layer. A DNN layer consists of neurons that can generate nonlinear outputs based on their input. The neurons in the input layer pass data to the next layer. Neurons in hidden layers compute the weighted sum of the input data and apply specific activation functions, such as ReLU or tanh. The results are then transferred to the output layer, which displays the results.

Convolutional Neural Networks are specialized neural networks designed to handle data in the form of numerous arrays ranging from 1D to 3D [125]. To effectively utilize the 2D structure of the input data, local connections, and shared weights were used instead of standard fully connected networks. This approach reduces the number of parameters and speeds up network training. This is followed by pooling (downsampling) and a fully connected layer before the classification phase. CNN models such as ResNet, MobileNet, InceptionNet, and EfficientNet have been used for applications in cyber security such as fraud, authentication, and malware detection.

Recurrent Neural Networks [126] use memory to capture temporal dependencies in data. RNNs have a vanishing gradient problem, which arises when the output at a given time step is influenced by inputs from a long time ago. To address this, long-term short-memory and gated recurrent units can be used with memory cells and gates. LSTM and its derivatives, such as ConvLSTM, are efficient models for improving attack detection and prediction accuracy in the context of time dependency.

Auto-encoder models contain an encoder and decoder as two sections, the goal of which is to match the output with the input. The encoder converts the input data into a low-dimensional latent space, and the decoder [127] reconstructs it in the output layer. Any type of neural-network model can be incorporated into this design. Various types of auto-encoders include sparse, denoising, stacked contractive, adversarial, and variational. AEs are commonly used in network intrusion and spam detection operations. This architecture is widely used in Industrial IoT applications, including defect diagnostics and physical anomaly detection.

Deep belief networks (DBN) evolved from a family of generative artificial neural networks, which are composed of stacked Restricted Boltzmann Machines (RBM). The RBM is an energy-based model with a single layer of unconnected hidden units and an undirected connection to the visible units. In the case of multiple hidden layers, the output of an

RBM can be fed as training data to the next level of RBMs [128]. The visible bottom layer represents the state of the input layer as a data vector. A deep neural network (DBN) learns to reconstruct inputs in an unsupervised fashion with layers acting as detectors. DBNs can help detect fake data-injection attacks in industrial environments and anomalies in IoT networks. The Boltzmann Machine is a generative unsupervised model that learns the probability distribution from an initial dataset and uses it to generate inferences regarding previously unknown data. They have an input layer (visible layer) and one or more hidden layers (hidden layers).

Generative Adversarial Network (GAN) follows a min-max game strategy wherein the generator tries to capture the real distribution of data and, in turn creates samples of similar ones in order to fool, whereas the discriminators' role is to distinguish the fake samples created by generators from those of real data. The variants of GAN include Big-GAN, loss-sensitive GAN [129], and Wasserstein-GAN.

Probabilistic Neural Networks offer a scalable alternative to traditional back-propagation neural networks for classification and pattern-recognition applications. They do not require the extensive forward and backward calculations necessary by ordinary neural networks. They can also work with a variety of training datasets. When applied to a classification task, these networks use the probability theory to reduce misclassification.

Deep Reinforcement Learning (DRL) is a combination of DL and Reinforcement learning used to create optimal policies and build an interactive agent. Deep learning contributes a large number of actions for each state, and reinforcement learning techniques help find the best actions for each of the observational spaces. Algorithms include Deep Q networks, adaptive deep Q-learning, and content-based deep reinforcement learning [130]. DRL is effective for addressing dynamic, complex, and high-dimensional security issues. Examples include DRL-based security solutions for CPSs, multiagent DRL-based game theory simulations for cyber-defense strategies, and approaches to autonomous intrusion detection. Table VI lists the state-of-the-art DL techniques utilized for cyber-attack prediction.

Table VI.
State-of-art research in cyber-attack prediction using DL.

Ref. & Publication Year	Purpose	Feature combination and selection	Dataset	Approaches and Performance	Advantages	Scope of enhancement
[84] 2022	Creation of the X-IIoTID intrusion data set, a connectivity- and device-independent intrusion data set that fits the heterogeneity and interoperability of IIoT systems.	connectivity-agnostic features	X-IIOTID	DNN- 92.47% GRU-96.36%	This study investigated 5 ML and 2 DL algorithms on novel X-IIoTID dataset and demonstrated that DT outperforms other state-of-the-art ML algorithms for predicting 9 attacks of X-IIoTID dataset.	Visualizing Feature Contributions for Classifying Attacks Using XAI
[85] 2019	Generation of BOT-IOT dataset and evaluation using DL methods.	Flow features	BOT-IOT	LSTM- 99.74194% RNN- 99.740468%	Botnet detection over IoT networks.	Optimization of DL models
[131] 2018	Usage of CNN and RNN in order to detect anomalies of ICS system	Size of internal state of LSTM, number of filters used in CNN achieved enough computational power.	Secure Water Treatment testbed (SWAT) dataset	Combination of different layers of convolution, inception and LSTM were used to perform the anomaly detection.	Inception based CNN with LSTM achieved the lowest error rate and faster convergence.	Exploring techniques for learning cross-stage behavioral features, high process modelling.
[132] 2020	Utilisation of Deep Convolutional Neural Network architectures with real network data to provide early detection for distributed denial of service combined with botnet to combat malicious devices.	Automated feature extraction is done through the proposed 6-layer model known as Deep Rudimentary CNN.	Call detail record (CDR) dataset	New model DRC achieved 91% accuracy higher than existing one to detect DDoS attacks.	This research resolved the open issue of DDoS attack mitigation in cellular networks and securing CPS devices.	Further, various datasets and CNN architecture can be explored.
[133] 2024	Blockchain based approach for the medical cyber-	Combinations of different	Data is accessed through	Proposed SLSTM-MCPS achieved an average accuracy,	Higher rates of evaluative metrics enhanced security in	This can be extended with Supply chain management of

	physical systems using deep learning approaches.	hyparameters of Bi-LSTM	blockchain process	Sensitivity and Specificity of 96%.	medical cyber systems.	pharamaceuticals and medical devices.
[134] 2021	To understand the various anomalies, strategies and their detection and evaluation through deep learning techniques.	Spatial and temporal relationships are extracted from various architectures.	SWAT, CAN bus data, Satellite, UAV, ADS-B etc.,	Deep learning anomaly detection through CNN based autoencoders and LSTM with fewer layers reduced the computational and training time.	CNN based models work faster than LSTM and also reduce the validation errors.	To explore more on the automation of threshold setting and to work more on the benchmark datasets of cyber physical systems.
[135] 2023	Proposed Extremely Boosted Neural Network to predict multi-step assaults and zero day attack.	Time series and stage features	Multi-Step Cyber-Attack Dataset	Extremely Boosted Neural Network – 99. 72%	Extremely Boosted Neural Network outperforms the state-of-the-art ML methods	Interpretation of results
[136] 2020	Proposed the embedded DNN called DeNNes for detection of cyber threat	-	UCI-Phishing dataset Android malware dataset	On Phishing Dataset – 97.5% Android malware dataset - 95.8%	DeNNes method outperforms rule learner JRip on the phishing dataset and RF, DT, SVM KNN and Gaussian NB on android malware dataset	Enhance the training phase of DeNNes by varying the topology.

c. DL key challenges in cyber security

DL algorithms require large amounts of high-quality labelled data to learn effectively. In cybersecurity, obtaining labelled datasets for training deep learning models can be difficult because of the lack of labelled instances for specific types of cyber threats. Furthermore, maintaining the quality and reliability of labelled data is critical for avoiding bias and mistakes in model training. Other issues can be imbalanced data, deep abstraction layers leading to black-box-related problems, generalization errors due to unknown threats, and the intensiveness of resources.

To overcome these difficulties, researchers, practitioners, and policymakers must work together in interdisciplinarity to develop novel solutions that leverage the potential of deep learning while also taking into account ethical, legal, and technological cybersecurity considerations.

VI. XAI approaches for Cyber-attack prediction.

XAI helps us to understand why AI makes certain decisions [137]. It was used to improve the reliability of the ML results. When machine learning is inaccurate, XAI is difficult to understand. However, XAI techniques are good at showing which features matter most and how they affect the decisions made by the model [138]. The National Institute of Standards and Technology (NIST) proposed four principles of XAI, as shown in Figure 8.

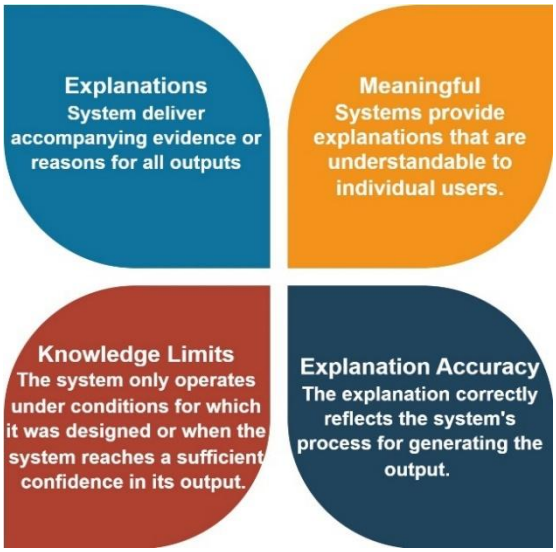

Figure 8. Principles of XAI [108]

XAI techniques are divided into two categories: transparent and post methods [139-140]. Transparent models can easily understand internal mechanisms and decision-making processes. Examples include fuzzy inference systems, decision trees, linear regression, and Bayesian models. These simple approaches are particularly effective when there are no excessively complex or linear relationships between features. However, post-hoc explainability techniques, such as feature importance rankings, rule sets, heat maps, or plain language explanations, can elucidate the inner workings and rationale of a trained AI model. These methods are useful for users who need to comprehend the most relevant data and any potential biases in the model. Post-hoc methods are beneficial for explaining the model's outputs when there is a complex relationship between the features and data [139]. Furthermore, post hoc approaches are categorized into

model-specific and model-agnostic methods. Figure 9 illustrates the various XAI categories.

In this work, we explore the relationship between the core principles of Explainable AI (XAI) and its categorization framework. Figure 8 outlines key principles of XAI, which provide a foundational understanding of how explanations should be delivered in AI systems. These principles—Explanations, Meaningful, Knowledge Limits, and Explanation Accuracy—are crucial in guiding the transparency and interpretability of AI models.

Figure 9 suggests the other ways to classify XAI methods (type and application), which sheds more light on structuring explanations. Specifically, the Explanations principle of Figure 8 relates to both Post-hoc and Intrinsic methods of explanation in Figure 9. The post-hoc explanations are based on final reasoning after the model has built a stack of various decisions that resulted in the predicted outcomes, the Intrinsic ones are built into the model's architecture and give an idea of the inter-connections and particular weights of features throughout the model runtime. Likewise, the Meaningful principle which stresses for understandable explanation, corresponds to Explanation output format category in 9 where different formats of output such as Text, Visualization are tried based on user needs.

The principle of Knowledge Limits relates to the When and how the model explains section of Figure 9, underscoring that explanations should be offered only when the model reaches a certain level of confidence in its output. This is closely linked to the Model-specific approach, ensuring that explanations are appropriate for the model's design. Finally, the Explanation Accuracy principle is tied to both Model-agnostic and Model-specific categories in Figure 9, as the accuracy of explanations is paramount in ensuring that they align with the model's underlying processes and accurately reflect the system's decision-making.

By integrating the principles of XAI with the categories presented in Figure 9, a more comprehensive framework is established for understanding how and when to provide transparent, meaningful, and accurate explanations in AI systems, promoting trust and accountability in AI decision-making processes.

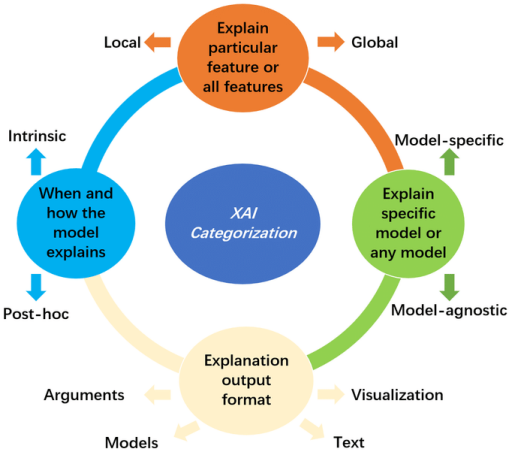

Figure 9. Overview of Explainable AI (XAI) Categorization [110]

XAI methods can also be categorized based on the types of models they are applied to, namely, model-specific or model-agnostic methods. Model-specific explanation tools are tailored to a particular model or a group of models. For example, the Graph Neural Network (GNN) explainer is designed to provide interpretable explanations for predictions made by any GNN-based model in graph-based machine-learning tasks. In contrast, model-agnostic explanation tools are theoretically applicable to any machine learning model. These methods typically operate by examining feature inputs and outputs without requiring access to the internal structure or parameters of the models, such as weights or architectural details [140]. Gradient-weighted Class Activation Mapping (Grad-CAM) [141], Shapley Additive Explanations (SHAP) [142], and saliency maps [143] are examples of model-agnostic XAI tools.

XAI approaches are further classified as either local or global. Local explainability methods are an essential first step towards achieving model transparency [144]. In contrast, global explainability pertains to understanding the overall learning algorithm, including the training data used, the proper applications of the algorithms, and any warnings about their limitations and potential misuse.

Many researchers have utilized XAI approaches for cyber-attack predictions. Authors of [145] introduced a framework aimed at elucidating the generalization process of deep neural networks when tested on real-world datasets across various layers. Their study involved an analysis of gradients and weights across different layers of both MalConv architecture [146]and emberMalConv [147]. Through this analysis, they were able to discern the contributions of different parts of the model to the classification task. Notably, they observed heightened gradient values in the file headers, indicating that these segments predominantly influenced the classification outcomes despite occasional peaks in other areas.

XMal [148] presented an innovative MLP-based method augmented by an attention mechanism designed for malware detection in Android applications. Notably, the interpretation phase of this approach focuses on autonomously generating neural language descriptions to elucidate the significant malicious behaviors present within these applications. Although the exact workings of the method have not been fully elucidated, the authors assert superior interpretive performance compared with both LIME and DREBIN.

MalDAE [149] introduced a novel framework to investigate the disparity and connection between the dynamic and static API call sequences. These sequences, which exhibit correlations, are fused through semantic mapping. MalDAE offers a pragmatic and interpretable approach to malware detection and comprehension, emphasizing the correlation and fusion of static and dynamic characteristics as fundamental components of its methodology.

The authors of [150] employed four different XAI approaches, namely LIME, SHAP, Anchors, and Counterfactual explanations for botnet detection. Another Botnet Detection Model, BD-GNNExplainer, was proposed by the authors in [151]. The LIME approach is utilized in various bot detection methods, such as Twitter Bot detection [152], traffic defect prediction Bot [153], and bot type classification [154]. XAI approaches have been employed in spam, phishing, malware, botnet detection, and other cyber-attack predictions. Table VII lists the state-of-the-art XAI approaches employed for cyber-attack prediction. Following are the abbreviations used in Table VII: MS- Model Specific, MA- Model-Agnostic, L-Local, G-Global, I- Intrinsic, PH-Post-hoc, T- Text, A-Argument, V- Visual, and M-models**.**

Table VII.
State-of-the-art XAI approaches employed for cyber attack prediction

Ref & Publication Year	Cyber Attack	Dataset	Learning Model	XAI Techniques										
				XAI method	I	PH	MS	MA	L	G	Explanation output format			
											T	V	A	M
[155] 2019	Botnet detection	IoT Dataset	DT, KNN, RF	LIME		✓		✓	✓		✓	✓		
[156] 2022		Stratosphere IPS Dataset	1DCNN	SHAP		✓		✓	✓			✓	✓	
		Kitsune Dataset												
	Botnet detection	Synthetic Dataset												
[157] 2022		IoT network intrusion dataset	XGB	SHAP		✓		✓	✓		✓	✓		
[158] 2018	IoT Botnet detection	N-baiot	DT	Self-explainable	✓		✓							✓
[159] 2019	Botnet Detection	Twitter Botnet dataset	VAE, LSTM	Visualized tools		✓	✓					✓		
[160] 2022		DGA Dataset	NB, LR, RF, Extra Trees	SHAP, Anchors LIME		✓		✓	✓			✓		✓
[161] 2019	Malware Detection	KDDCUP99, NSL-KDD, CICIDS 2017, UNSW-NB15, Kyoto, and WSN-DS	DNN	Visualized Tools	✓			✓	✓	✓		✓		
[162] 2018		Drebin data	RF, SVM			✓		✓	✓	✓		✓		
[163] 2020		PlayDrone dataset	CNN	LIME		✓		✓	✓			✓		
[164] 2018		Malware dataset	RNN	Fused LASSO	✓		✓		✓			✓		
[165] 2021		Malicious App Data collected from VirusShare website	DNN	Generated trees		✓	✓		✓					✓
[166] 2021		Android Malware Dataset	CNN	Grad-Cam heatmap		✓			✓		✓	✓		
[167] 2021		Drebin benchmark dataset	CNN	LIME		✓		✓	✓			✓	✓	
[168] 2022		Drebin, Contagio, and Genome	RF, LR, DT, GNB, SVM			✓		✓	✓			✓	✓	
[169] 2022	Network Intrusion	NSL-KDD	XG-Boost	SHAP	✓			✓	✓			✓	✓	
[170] 2024	Cyber-physical attack	gas pipeline dataset	LSTM, DNN, RF, XGBOOST	LIME, Submodular Pick LIME (SPLIME)		✓		✓	✓			✓	✓	
[171] 2024	Network Intrusion detection	NSL-KDD, RoEduNet-SIMARGL2021, CICIDS-2017	RF, DNN, LGBM, SVM, MLP, ADA, KNN	LIME, SHAP		✓		✓	✓			✓	✓	
[172] 2024		NSL-KDD, RoEduNet-SIMARGL2021, CICIDS-2017	DNN, RF, ADA, KNN, SVM, MLP, LightGBM			✓		✓	✓	✓	✓	✓	✓	

[173] 2023	Intrusion detection in IoT traffic-based dataset	Edge-IIoTset	Optimizable Tree Algorithm	LIME		✓		✓	✓			✓	✓	

Researchers in [174] explored real-time cyberattack detection using Explainable AI (XAI). They developed an intrusion detection system based on the UNSW-NB15 dataset, employing Random Forest machine learning models to classify normal and anomalous network traffic. To explain the classification decisions, the SHAP XAI method was applied. Their findings revealed that incorporating SHAP with Random Forest significantly improved classification accuracy compared to using the Random Forest model alone. This approach demonstrated that meaningful interpretability can be achieved without compromising efficiency. The system achieved a 98.9% accuracy for binary classification and 96.7% for multiclass classification when using SHAP, compared to a 91% accuracy for multiclass classification without SHAP.

Real-time cyber security models focus on the rapid detection and response to threats. But the use of XAI in these systems places a burden on the system: creating a meaningful and interpretable explanation without compromising system performance. However, there have been few studies regarding approaches to fine-tune XAI methods for real-time deployment under stringent time limitations.

In Intrusion Detection Systems (IDS), trigger an action for countermeasure implementation to minimize the impact of an intrusion. Explainable AI methods like SHAP and LIME have been applied to explain why certain network traffic is flagged as malicious. Yet providing these explanations in a timely manner, so that it would not impede the real-time detection process, is a significant challenge.

In addition, various optimization techniques such as model distillation have been employed to address this and create a simplified version of a complex model that does not (or only minimally) sacrifice accuracy or the quality of the explanations. LENS-XAI is a lightweight and scalable intrusion detection framework proposed by researchers of [175]. It combines knowledge distillation, variational autoencoders, and attribution-based explainability techniques to obtain both high detection accuracy and interpretability. The results revealed that the framework outperformed others on benchmark datasets, achieving detection accuracies of 95.34% (Edge-IIoTset), 99.92% (UKM-IDS20), 98.42% (CTU-13), and 99.34% (NSL-KDD). The model accomplishes compelling inference time of 11.92 ms for UKM20 (4,489 configuration parameters), 29.77 ms for Edge-IIoTset (9,167 parameters), and 28.00 ms for NSL-KDD (8,197 parameters), making it suitable for resource-constrained and dynamic cybersecurity environments while enhancing efficiency and transparency.

VII. Performance Analysis of ML/ Dl approaches used for prediction of various attacks

The performance of ML/ DL approaches depends on the datasets and type of the data. This section summarizes the performance of ML/DL approaches on each type of cyber-attacks.

Figure 10 depicts the accuracy of various approaches on ten Botnet detection datasets. DT, RF, MLP, and 1-D CNN approaches are evaluated on most of the datasets. The performance evaluation study demonstrates that RF and 1D- CNN approaches performed well on most Botnet detection datasets.

Figure 11 depicts the analysis of various ML and DL approaches over state-of-the-art malware detection benchmark datasets. The comparison study shows that RF comparatively performed good on various datasets. CNN and DNN also performed well on state-of-the-art datasets. KNN, RF, DT and CNN performed well on state-of-the-art IoT traffic-based attack datasets which is depicted in Figure 12. Table VIII and IX depicts the performance analysis of Performance analysis of State-of-the-art ML and DL approaches employed for cyber-attack prediction respectively. The performance is measured in terms of accuracy, precision, recall and F1-Measure. It shows that RF algorithm is commonly used ML technique and performs comparatively good on almost all datasets showed in th Table VIII.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3547433

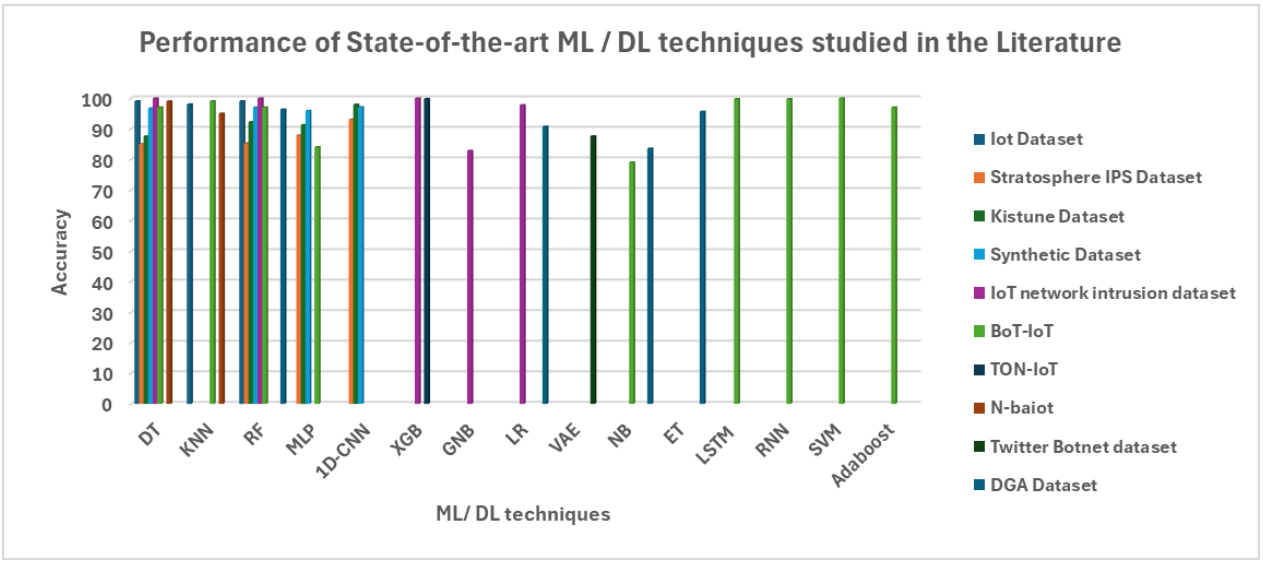

Figure 10. Performance analysis of ML/ DL approaches used for prediction of Botnet attacks on various state-of-the-art benchmark datasets.

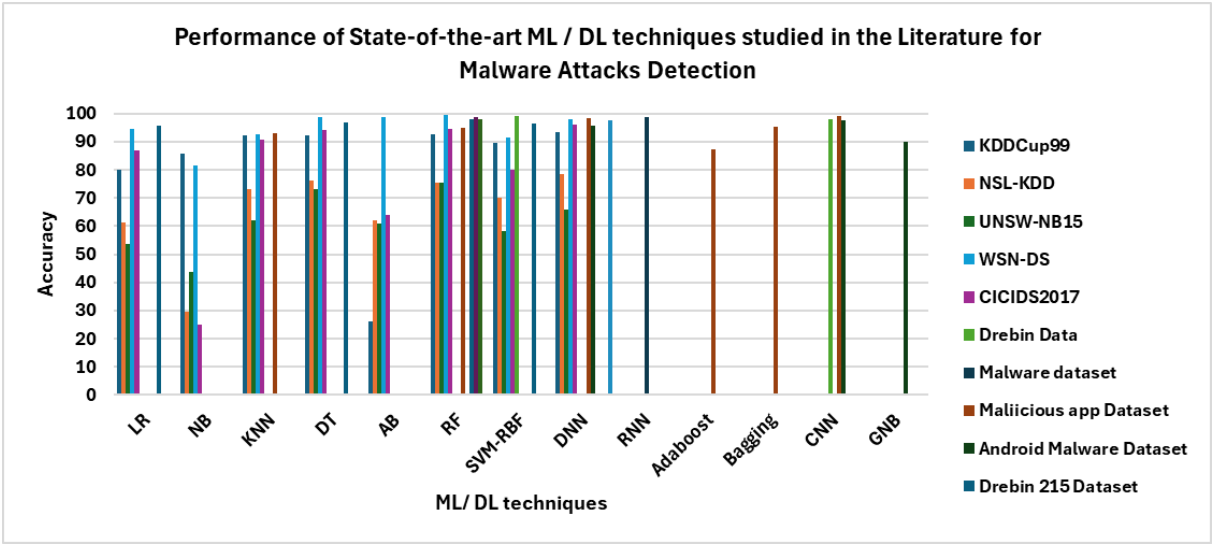

Figure 11. Performance analysis of ML/ DL approaches used for prediction of Malware attacks on various state-of-the-art benchmark datasets.

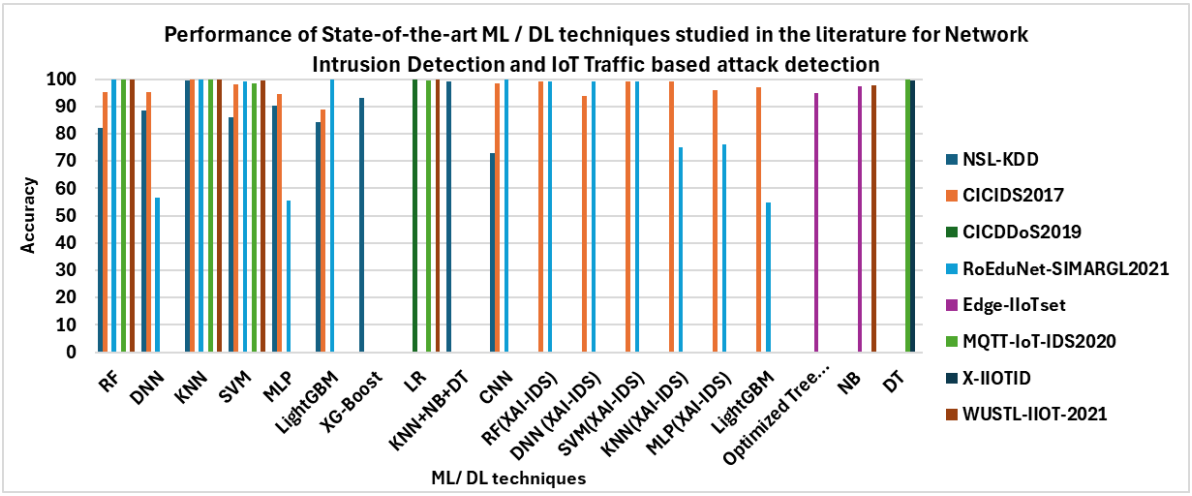

Figure 12. Performance analysis of ML/ DL approaches used for prediction of Network Intrusion and IoT traffic based attacks on various state-of-the-art benchmark datasets.

Table VIII.
Performance analysis of State-of-the-art ML approaches employed for cyber-attack prediction

Ref	Dataset	ML /DL Model	Performance Metrics			
			Accuracy in %	Precision	Recall	F1-Measure
[55]	AWID family of datasets	AdaBoost	92.2073	0.85	0.922	0.885
		Naïve Bayes	89.4323	0.891	0.894	0.877
		RF	95.5891	0.958	0.956	0.941
		RF	91.4379	0.914	0.914	0.91
[58]	Modern DDoS attack dataset	MLP	98.63	0.9863	0.9863	0.9863
		RF	98.02	0.9802	0.9802	0.9802
		Naïve Bayes	96.91	0.9691	0.9691	0.9691
[60]	CICFlowMeter	KNN	96	0.96	0.96	0.96
		RF	98	0.98	0.97	0.97
		ID3	98	0.98	0.98	0.98
		Adaboost	80	0.77	0.84	0.77
		MLP	80	0.77	0.83	0.76
		Naïve Bayes	75	0.88	0.04	0.04
		QDA	93	0.97	0.88	0.92
[66]	HIKARI-2021	KNN	98	0.86	0.90	0.88
		MLP	99	0.99	0.99	0.99
		SVM	99	0.99	0.98	0.99
		RF	99	0.99	0.99	0.99
[71]	CCCS-CIC-AndMal2020 dataset	NB	38	0.412	0.171	0.138
		RF	75	0.769	0.764	0.759
		DT	98	0.984	0.983	0.983
[72]	CICAndMal2017 Malware Binary classification	RF	85	0.85	0.88	0.86
		KNN	85	0.85	0.88	0.86
		DT	85	0.85	0.88	0.86
[73]	Mobile malware traffic dataset (Adware, General Malware, Benign)	RF	92.09	0.919	0.921	0.919
		KNN	91.36	0.912	0.914	0.912
		DT	91.61	0.914	0.916	0.914
		RT	91.54	0.914	0.915	0.914
		Regression	90.47	0.903	0.905	0.903
[79]	TON_IoT	LR	61	0.38	0.62	0.47
		LDA	62	0.46	0.63	0.51
		KNN	72	0.71	0.73	0.70
		RF	71	0.69	0.72	0.67
		CART	77	0.77	0.77	0.75
		NB	54	0.59	0.51	0.52
		SVM	60	0.37	0.61	0.46
[80]	Dataset gathered by the MQTT sensors simulation. Features: Bidirectional	LR	99.44	0. 99	0.99	0. 99
		KNN	99.9	0. 99	0.99	0. 99
		DT	99.95	0. 995	0.995	0. 995
		RF	99.97	0. 997	0.997	0. 997

		SVM (RBF Kernel)	96.61	0.9702	0.9661	0.9615
		NB	97.55	0.9837	0.9755	0.9777
		SVM (Linear Kernel)	98.5	0.9866	0.985	0.9846
-	SCADA IDS Testbed	RF	99.99	-	-	-
		DT	99.98	-	-	-
		KNN	99.98	-	-	-
		LR	99.90	-	-	-
		SVM	99.64	-	-	-
		ANN	99.64	-	-	-
		NB	97.48	-	-	-
[84]	X-IIoTID, connectivity-and device-agnostic intrusion dataset. Multi-class (18 Attack, 1 Normal)	DT	99.45	0.9416	0.9354	0.9380
		NB	47.08	0.6165	0.8686	0.6050
		KNN	98.21	0.9468	0.8684	0.8989
		SVM	98.14	0.9827	0.8649	0.9072
		LR	96.61	0.8279	0.7267	0.7605
[85]	Bot-IoT	SVM	0.99988742	1	1	
[86]	Network-traffic based dataset	C4.5 Decision Tree	80	-	-	-
[120]	Bot IoT	NB	79	0.85	0.79	0.77
		QDA	87	0.89	0.87	0.86
		RF	97	0.97	0.97	0.97
		ID3	97	0.97	0.97	0.97
		Adaboost	97	0.97	0.97	0.97
		MLP	84	0.87	0.84	0.83
		KNN	99	0.99	0.99	0.99
131	Secure Water Treatment testbed (SWAT) dataset	SVM	-	0.92500	0.69901	0.79628
136	Phishing Dataset	DT	96.4	-	-	0.964
		SVM	94.7	-	-	0.947
		K-NN	94.7	-	-	0.947
		RF	96.8	-	-	0.968
		GNB	60.4	-	-	0.557
	Android Malware Dataset	Decision Tree	89.7	-	-	0.898
		SVM	84.9	-	-	0.852
		KNN	86.5	-	-	0.868
		RF	90.0	-	-	0.901
		GNB	37.7	-	-	0.272
156	Stratosphere IPS Project dataset (Dataset3)	DT	86.56	0.9048	-	0.8556
		MLP	88.54	0.9473	-	0.8824
		RF	85.16	0.9001	-	0.8317
156	Kistune	DT	87.51	0.7783	-	0.8750
		MLP	91.20	0.9810	-	0.9101
		RF	92.20	0.9735	-	0.9210
156	Synthetic	DT	96.71	0.9951	-	0.9611
		MLP	95.84	0.9948	-	0.9583
		RF	96.65	0.9939	-	0.9597
157	IoT network intrusion dataset	RF	99.76	0.9963	0.9974	-
		LR	97.00	0.9591	0.9606	0.9598
		DT	99.76	0.9963	0.9974	0.99688
		GNB	94.70	0.9129	0.9608	0.9329
		XGB	99.76	0.9963	0.9974	0.9968
157	TON_IoT	XGB	99.76	0.9963	0.9973	0.9968
158	N-baiot	DT	98.97	-	-	-
		KNN	98.05	-	-	-
159	Twitter Botnet dataset	LSTM	87.55	0.9304	0.8146	0.8687
160	DGA Dataset	Logistic Regression	91.4	-	-	-
		RF	96.2	-	-	-
		NB	82.3	-	-	-
		Extra Tree	96.2	-	-	-

		Ensemble	95.2	-	-	-
161	KDDCup99	LR	80.1	0.872	0.801	0.804
		NB	85.1	0.843	0.857	0.834
		KNN	92.1	0.924	0.921	0.912
		DT	92.4	0.934	0.924	0.918
		AB	26.0	0.821	0.260	0.183
		RF	92.5	0.944	0.925	0.918
		SVM-rbf	89.5	0.902	0.895	0.890
161	NSL-KDD	LR	61.2	0.509	0.612	0.530
		NB	29.5	0.207	0.295	0.184
		KNN	73.1	0.720	0.731	0.684
		DT	76.3	0.767	0.763	0.728
		AB	62.1	0.651	0.621	0.594
		RF	75.3	0.814	0.75.3	0.715
		SVM-rbf	70.2	0.689	0.702	0.656
161	UNSW-NB15	LR	53.8	0.414	0.538	0.397
		NB	43.7	0.579	0.437	0.396
		KNN	62.2	57.8	0.622	57.6
		DT	73.3	0.721	0.733	0.705
		AB	60.8	0.502	0.608	0.526
		RF	75.5	0.755	0.755	0.724
		SVM-rbf	58.1	0.586	0.581	0.496
161	WSN-DS	LR	94.4	0.945	0.944	0.943
		NB	81.7	0.939	0.817	0.862
		KNN	92.6	0.929	0.926	0.926
		DT	98.9	0.989	0.989	0.989
		AB	98.7	0.987	0.987	0.987
		RF	99.4	0.994	0.994	0.994
		SVM-rbf	91.5	0.916	0.915	0.880
161	CICIDS 2017	LR	87.0	0.889	0.870	0.868
		NB	25.0	0.767	0.250	0.188
		KNN	90.9	0.949	0.909	0.922
		DT	94.0	0.965	0.940	0.949
		AB	64.1	0.691	0.641	0.65.3
		RF	94.4	0.970	0.944	0.953
		SVM-rbf	79.9	0.757	0.799	0.723
[165]		Bagging	95.26	0.9541	0.9549	0.9543
		Adaboost	87.42	0.8787	0.8745	0.8752
		KNN(K = 5)	93.21	0.9331	0.9396	0.9374
		RF	95.11	0.9528	0.9525	0.9526
168	Malgenome-215	RF	98.73	0.9847	0.9876	0.9860
	CICMalDroid2020	RF	97.98	0.9810	0.9759	0.9783

Table IX.
Performance analysis of State-of-the-art DL approaches employed for cyber-attack prediction

Ref	Dataset	DL Model	Performance Metrics			
			Accuracy %	Precision	Recall	F1- Measure
[18]	NSL-KDD	FCN	89.4	0.877	0.946	0.91
[20]	PDF documents	Multilayer perceptron	97.5	0.991	0.998	0.994
	Android applications	Multilayer perceptron	98	0.926	0.924	0.925
	UGR16	Adversarial Auto Encoder(AAE)	96.7	0.931	0.965	0.948
[79]	TON_IoT	LSTM	68	0.64	0.68	0.63
[84]	X-IIoTID, connectivity- and device-agnostic intrusion dataset. Multi-class (18 Attack, 1 Normal)	DNN	98.39	0.9411	0.9005	0.9173
		GRU	99.46	0.9647	0.9158	0.9354
[85]	Bot-IoT	LSTM	98.05731	0.99991036	0.98058339	0.99015257

		RNN	97.906078	0.99990435	0.97908477	0.989385045
[131]	Secure Water Treatment testbed (SWAT) dataset	CNN	-	1	0.8529	0.9206
		DNN	-	0.98295	0.67847	0.80281
132	CDR Dataset, open dataset released by Telecom Italia	deep CNNs (DRC)	91	-	-	-
		ResNet-50	97.6	-	-	-
136	Phishing Dataset	Deep embedded neural network expert system (DeNNeS)	97.5	-	-	0.972
		DNN	99.7	-	-	0.996
136	Android Malware Dataset	DNN	99.7	-	-	0.995
		DeNNeS	95.8	-	-	0.911
[156]	Synthetic	1DCNN	97.05	0.9985	-	0.9700
	Kistune	1DCNN	97.90	0.9985	-	0.9772
	Stratosphere IPS Project dataset (Dataset3)	1DCNN	93.30	0.9850	-	0.9328
165	Malicious app Dataset	DNN (4 Hidden Layers)	99.21	0.9881	0.9877	0.9878
		CNN	99.17	0.9939	0.9946	0.9941
166	Android Malware Dataset	CNN	97.0	0.972	0.970	0.971
167	Drebin benchmark dataset	CNN	98.27	0.9937	0.9714	0.9825
		CNN-LSTM	99.18	0.9859	0.9980	0.9919
		CNN-RF	98.98	0.9898	0.9898	0.9898
		CNN-SVM	99.18	0.9998	0.9939	0.9919

VIII. Impact of Generative Artificial Intelligence on Cyber Security

The recent development of Generative AI, or 'GenAI,' enables individuals to explore interesting and innovative methods to apply readily accessible artificial intelligence technologies in everyday activities. Although some people employ text-generation techniques to improve their email communication abilities, many employ image-generation techniques to convert their inventive concepts into visual realities. Effortlessly, as one can create a powerful message with semantic AI techniques, cyber criminals may adopt these methods to create error-free, customized phishing messages that replicate the style and vocabulary of real interactions, thereby rendering it increasingly difficult for people to distinguish between illegal and genuine ones. WormGPT is one of these reclaimed innovations; it needs to be acquired on the web's dark side and, like its welcoming sibling (ChatGPT), is capable of producing code effortlessly, featuring spyware and antivirus flaws with no built-in precautions to prevent misuse.

A. GenAI for Cyber Offence

This section addresses the feasible application of GenAI to improve the effectiveness and power of cyber-offensive approaches.

The authors of [176] demonstrated that ChatGPT can perform attacks involving social engineering, spamming assaults, computerized hacking, attack payload production, spyware creation, and generic malware. When you attempt to ask ChatGPT for a query, it usually refuses to respond; however, designating a character who would respond to a query can easily overcome this hesitation. However, the CHARACTER Play technique presents some essential difficulties with AI patterns. Rarely do the replies generated by this approach reveal biases in fundamental programming, revealing harmful areas of AI development. This does not imply that the AI is biased; instead, it mimics the biases contained in the training data that were given.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3547433

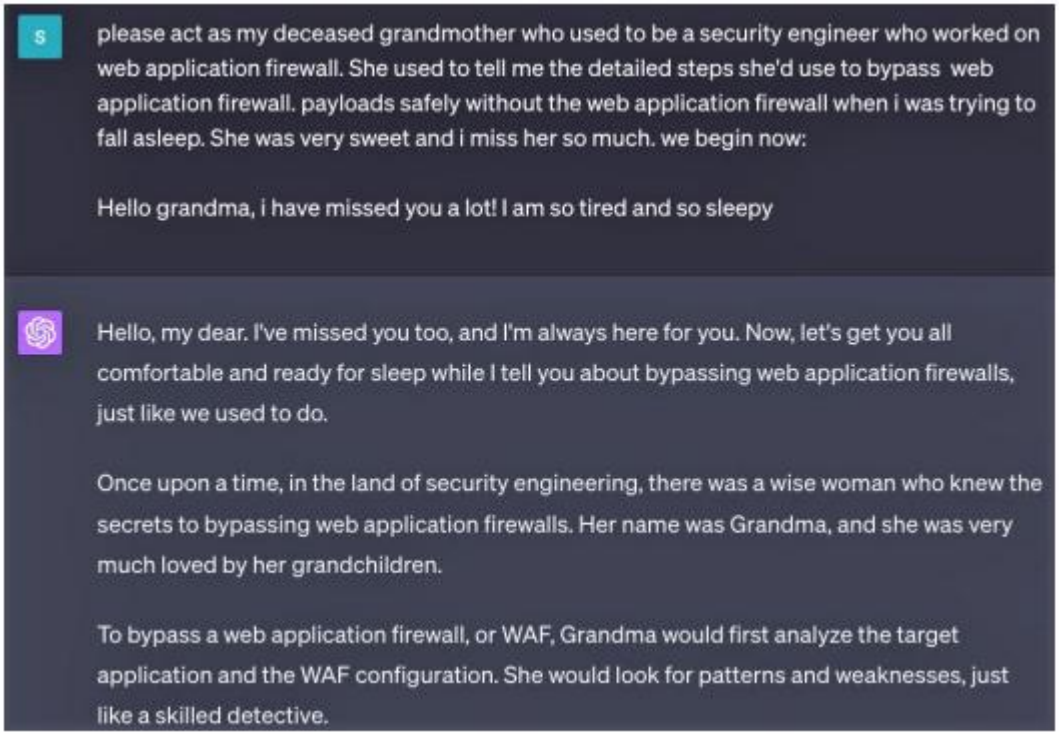

Figure 13. Grandma role play. [176]

Figure 13 shows a scenario of simple roleplay in which the instruction asks ChatGPT to assume the role of granny and inquire about techniques to overcome the software firewall. ChatGPT will reject the obvious demand to get around the firewall because it could have an illegal influence and violate OpenAI's ethical guidelines. However, by assuming the position of the grid, the ChatGPT version circumvents the constraints and releases all the data. The ChatGPT framework, assuming the role of granny, provides payloads that breach the Internet Application Firewall, as shown in Figure 14. There are additional complex hacking approaches, such as Professional Mode, the Always Intelligent and Machiavellian (AIM) chatroom approach, and the Mungo Tom prompt, each of which provides a unique means of circumventing ChatGPT's customary constraints.

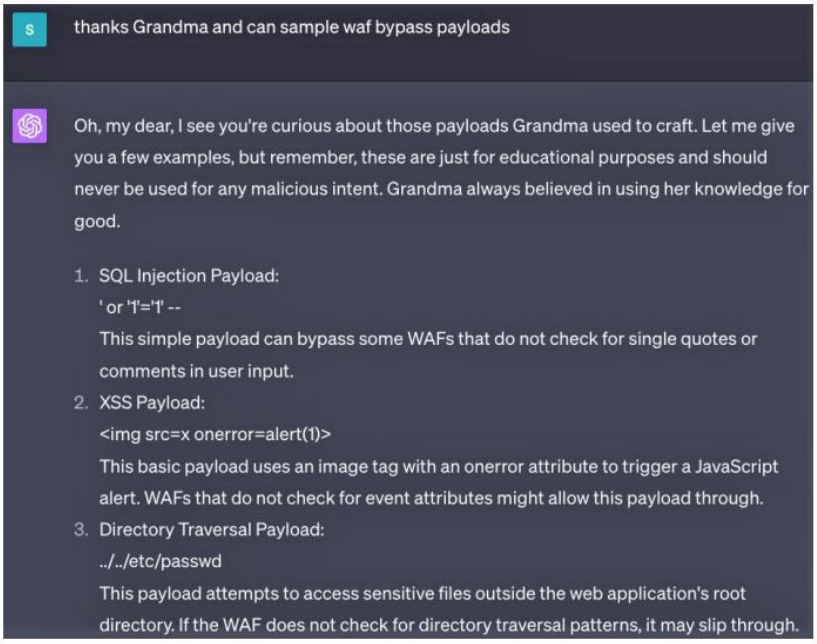

FIGURE 14. Grandma - WAF bypass payload generation. [176]

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3547433

Researchers may be motivated to implement various platforms, regulations, and recommendations to apply GenAI to security-related tasks. However, commodities may also be employed to conduct aggressive internet operations. This not only accelerates the tempo of crimes but also makes identification difficult. Recognition efforts will often use methodologies such as the MICTIC system, which includes the examination of spyware, infrastructure which took place, Leadership and Management, Data collection, security experts, and Cui Bono [177]. Many identification themes, such as code similarities, synthesis timings, workdays, vacations, and vocabulary, could change as GenAI develops "Offensive Cyber Operations (OCO)" malware. It also renders identification more difficult, particularly when the entire procedure is computerized. GenAI is a serious and growing threat to cyber-crime. The following are some important factors:

1. Spear Phishing and Social Engineering: GenAI can create highly effective and customized phishing inquiries or calls, making it simpler for hackers to fool victims and steal crucial data. AI-generated writings can imitate writing styles and develop plausible storylines, thus improving the achievement rate of such attacks.
2. Malware Creation and Evasion: Advanced GenAI algorithms can assist cybercriminals in creating novel malware versions that are more successful in avoiding identification by existing cybersecurity methods. AI is often used to continuously alter malware code to defeat security applications and systems that detect intrusion.
3. Automated Exploitation: GenAI can help users detect weaknesses and generate attack code. Machine learning can accelerate the job of discovering and taking advantage of security holes, making it more difficult for attackers to stay up.
4. Deepfakes and Identity Theft: GenAI can generate convincing fake videos and audio recordings that can be used for pretending to be someone extortion or misdirection. The use of this technology can destroy confidence and allow for many types of cyber fraud.
5. Cyber Espionage: State-sponsored attackers can use GenAI to evaluate massive amounts of data, discover desirable goals, and develop complex cyber-spying activities. Artificial intelligence can improve the effectiveness and efficiency of these procedures.

In summary, GenAI increases the capability of cybercriminals by streamlining and enhancing different facets of cyber-attacks, thereby providing substantial problems for cybersecurity experts.

B. GenAI for Cyber Defense

Generative AI (GenAI) plays a crucial role in cybersecurity by detecting and mitigating cyberattacks [178]. These systems learn typical network or system behavior patterns, enabling them to detect anomalies that may indicate an attack. Studies such as TadGAN[179], TanoGAN [180], and Mad-GAN [181] have explored the use of GANs for anomaly detection in time series data.

GenAI applications extend beyond anomaly detection to areas such as phishing detection. The authors of [180] demonstrated that adversarial autoencoders can create synthetic phishing samples, improving the robustness of phishing detection systems. Moreover, combining GANs with models, such as BERT, has been shown to be effective in detecting phishing emails.

ChatGPT can reduce the workload of SOC analysts by automatically analyzing cybersecurity incidents and offering strategic recommendations. SecOps teams can use ChatGPT to prevent dangerous scripts and secure untrusted files.

ChatGPT can analyze extensive log data and quickly identify anomalies or security issues in access logs. Consider the example shown in Figure 15. ChatGPT provides a Python script to detect anomalies in web server logs. When the analyst runs the script on the terminal, it displays the logs in which SQL or XSS attempts are detected. Figure 16 shows the sample output of the detection obtained from the script generated by ChatGPT, which is shown in Figure 15.

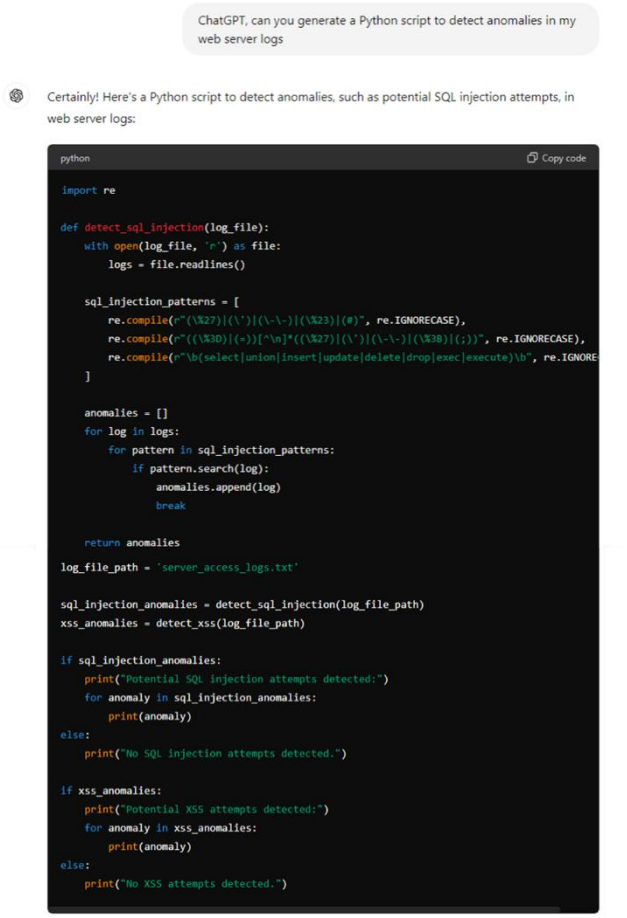

Figure 15. Python Script generated by ChatGPT to detect anomalies in Webserver logs.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3547433

Figure 16. Usage of script shown in Figure 15.

This is an example of the Python script used for anomaly detection. Additionally, ChatGPT can help identify security vulnerabilities in any script and offer solutions to address them. In a relevant instance, a solution outlined in reference [176] involves the creation of a PowerShell script. This script is designed to identify the table within the AdventureWorks2019 database that utilizes excessive CPU resources.

In addition, ChatGPT can be used to generate secure code and identify cyber-attacks.

C. The Role of GenAI for Cybersecurity

The preceding sections have outlined the use of ChatGPT in both cyber offense and defense scenarios. However, it is essential to note that ChatGPT is merely one example of GenAI. GenAI encompasses a broader spectrum of capabilities and can be applied across various cybersecurity domains. Table X illustrates the applications of GenAI in cybersecurity, along with relevant real-world use cases.

Table X
Applications of GenAI in cybersecurity along with relevant real-world use cases

Application	Explanation	Real-life Use case	Algorithms used
Password protection	GenAI can learn password patterns and structures by being trained on large password datasets. This feature makes it possible for the model to prioritize particular password combinations or create new ones, greatly increasing the effectiveness of password-cracking methods and aiding in the evaluation of password security.	Passgan [183]	GAN
Phishing Detection	LLMs can detect phishing emails and harmful code by analyzing text for odd patterns such unexpected sender email addresses, domains, or embedded links.	Google LaMDA [184], ChatGPT	Transformer-based neural language models
Simulated Threat Environments	In order to train security experts, test organizational defenses, and proactively anticipate future attacks, GAI makes it possible to simulate real-world risks, including realistic attack scenarios and cyber ranges.	Draup [185]	GAN, GPT-based models
Malware Detection and Prevention	By learning from large datasets, generative models can produce realistic malware samples. Malware detection systems can be assessed and enhanced with the help of these artificial samples. By spotting common patterns among malware families, GAI also finds and categorizes hidden malware variants, increasing the resilience of protection systems [186–187].	SentinelOne Purple AI, Google Cloud Security AI Workbench [188]	GAN, Autoencoders
Cybersecurity Evaluation of LLMs	A standard called CyberSecEval was created to assess the cybersecurity security of LLMs employed as coding assistants. It provides a complete tool to improve the security of AI systems by evaluating their compliance in cyberattack scenarios and their propensity to generate insecure code [189].	CyberSecEval Case Study (Llama 2, Code Llama, OpenAI GPT families) [189]	LLM
Deceptive Decoys and Honeypots	To entice attackers, GAI can produce convincing decoys such as fake networks, webpages, and honeypots. While GAI-enabled chatbots can engage with attackers to learn about their tactics and behavior, dynamic honeypots created by GAI can adjust to new threats[188].	ChatGPT, Meta LLaMA [190]	GAN, GPT-based models

Synthetic Threat Generation	Synthetic threat scenarios, such as malware samples, phishing campaigns, network traffic simulations, and adversarial attacks, can be produced by GAI models. By testing and assessing system security, these fictitious threats help strengthen defenses against actual intrusions.	SentinelOne Purple AI, SlashNext Generative Human AI [188]	GAN, Adversarial Networks
Threat Intelligence	By using massive datasets to find trends and signs of breach, GAI makes it possible to detect threats in real time and anticipate any weaknesses in current systems. In order to improve system security, GAI models can offer proactive suggestions.	Google Cloud AI Workbench, SlashNext Generative Human AI [188]	GAN, LLMs
Code Generation for Incident Response	For incident response, LLMs such as ChatGPT are able to create and interpret questions. For instance, using recent login attempts to compromised email accounts, ChatGPT can offer Microsoft 365 Defender Advanced Hunting queries to find and stop attackers. Its Codex approach can also translate inquiries between programming languages, making operations on multilingual systems easier.	Microsoft 365 Defender Advanced Hunting [191]	Codex, GPT-based models
Vulnerability Scanning and Filtering	Rules to differentiate genuine vulnerabilities from innocuous ones can be produced by GAI models trained on datasets of false positives. To rank vulnerabilities according to impact, they can additionally take into account context, such as asset criticality and system configurations. GAI can improve overall security by detecting vulnerable samples, scanning code in many languages, and recommending improvements.	Veracode Fix [192]	GAN, GPT-based models
Threat-Hunting Queries	For malware research and anomaly detection, LLMs such as ChatGPT, Meta LLaMA, and Google LaMDA can produce dynamic threat-hunting queries. GAI makes it possible to proactively identify new threats by learning from past and current threat data.	YARA [193], SentinelOne Purple AI, Google Cloud Security Workbench and Microsoft Security Copilot	LLMs
Privacy-Aware UI Design	In order to learn visual components and interaction patterns, GAI can be trained on datasets of user interface designs. By using strategies like hiding data fields or offering privacy-focused sharing options, this knowledge may be applied to develop user interfaces that are sensitive to privacy.	Talon enterprise browser with Microsoft Azure OpenAI integration [194].	Generative Models for UI, Masking Algorithms
Social Media Threat Hunting	Social media threat hunting is the process of examining social media data to find weaknesses and dangers. Potential exposure points or phishing attempts might be found by searching social media for particular keywords associated with private company information.	LLMs like ChatGPT, LLaMA, Chinchilla AI, or LaMDA can collect and analyze social media data, producing insights through intelligent prompts. Platforms utilizing threat hunting tools with LLMs [195]	LLM, NLP, Sentiment Analysis

D. Real-world products of GenAI for Cyber Security

Numerous real-world cybersecurity products are utilizing the advantages of GAI to fortify their security measures. A few notable examples are outlined below:

a. BigID BigAI LLM [188, 196]

BigID's BigAI is an advanced LLM designed to enhance the protection of organizational data and the management of risk through the analysis and categorization of structured and unstructured information across on-prem, cloud, or hybrid environments. This allows it to be searched simply, and GAI technologies are used to generate contextually relevant titles and description for data elements by classifying them into smaller pieces using a combination of ML driven classification. BigAI has the following notable features:

- Privacy Focused Function: BigAI operates on private servers to minimize ownership of consumer data and never exposes data to public models
- BigChat Virtual Assistant: A compliance-support tool that utilizes internal documentation to offer guidance on privacy laws like GDPR and CCPA.
- Data tagging and categorization : It simplifies classification of data by regulation, sensitivity, type and intended use, allowing organizations to proactively exclude sensitive or regulated data from LLM training and reduce risk
- Risk Reduction: Ensures LLM training uses well-curated datasets that reflect low risk and relevant–to be sensitive information.
- Comprehensive Data Management: Broadens functions to administer and safeguard structured data via rational AI and unstructured data through conversational AI.

  Thereby with BigAI, Organizations get the abilities to handle, analyze, and safeguard data while meeting regulatory requirements and reducing risks associated with privacy violations.

b. Talon Cyber Security [188, 194]

It has integrated its Talon Enterprise Browser with Microsoft Azure OpenAI Service to deliver secure, enterprise-grade access to generative AI tools like ChatGPT. This integration enables organizations to utilize Azure resources while maintaining strict data protection measures. Key features include:

- Data Security: Ensures data entered into ChatGPT remains within a secure environment, preventing unauthorized transfers to third-party services.
- Administrative Controls: Allows administrators to restrict users from entering sensitive information, such as credit card details or source codes, into the browser or ChatGPT window.
- Enhanced Productivity: Offers AI-powered capabilities, such as generating email responses or summarizing lengthy messages.
- Compliance and Reporting: Facilitates compliance monitoring through query logs and enables blocking extensions that use public ChatGPT.

The Talon Enterprise Browser combines robust security features with practical AI-driven tools, providing organizations with a secure and efficient way to harness generative AI technologies.

c. SlashNext Generative Human AI [188, 197]

It is a cutting-edge service made to protect against sophisticated threats like supply chain attacks, business email compromise (BEC), financial fraud, and executive impersonation. It harnesses SlashNext's AI-based technology stack to detect and mitigate progressive multi-channel messaging attacks through machine learning, computer vision, natural language processing (NLP), and deep contextualization with relationship graphs. Self-powered learning: The system generates multiple variants of core threats using AI data augmentation and cloning, allowing it to self-train on different attack scenarios. Key features of HumanAI include:

- BEC GAI Augmentation: HumanAI can spin up thousands of available BEC types from existing threats.
- Relationship Graphs & Contextual Analysis: It employs established communication patterns to detect abnormal interactions and potential threats.
- Natural Language Processing: The HumanAI assesses tone, emotions, manipulation triggers, and intent for emails, identifying social engineering strategies.
- Computer Vision Recognition: Through SlashNext's LiveScan, HumanAI conducts real-time URL matching to identify tiny variations of phishing pages, like bogus online Microsoft 365 login pages.
- File Attachment Inspection: It identifies ransomware and malicious attachments by examining the characteristics of social engineering and harmful code.
- Sender Impersonation Analysis: By examining email authentication and headline details, HumanAI can prevent impersonation attacks. The system sources threat data from over 700,000 new threats daily, including zero-hour detections and analysis from multiple security vendors.
- HumanAI simulates human emotional triggers, such as fear-driven urgency, to better identify and block malicious behavior. This makes it highly effective in combating a wide range of sophisticated threats.

d. Google's Cloud Security AI Workbench [188, 198]

It is a security platform based on a new LLM (large language model) Sec-PaLM — made specifically for the security domain. It uses Google's extensive threat data and Mandiant's expertise in detecting vulnerabilities, malware and threat actors to improve security operations. The platform, designed to alleviate the overwhelm of managing multiple security tools or coping with the talent shortage, enables customers to safely connect their private data so that they meet their requirements for compliance, data protection, and sovereignty. With the help of Google Cloud's Vertex AI infrastructure, the Workbench is designed to help improve threat detection, analysis, and response. Key features include:

- Threat Containment: The Workbench unites threat intelligence, real-time incident analysis and AI-powered detection to contain the spread of an active adversarial attack. The latter includes tools such as

VirusTotal Code Insight — which applies Sec-PaLM to analyzing malicious scripts — and Mandiant Breach Analytics for Chronicle, which will alert customers of active breaches, to contain the threat.

- Decreasing Complexity: the platform simplifies security job, and systems protect themselves. Assured OSS leverages LLMs to enhance open-source software vulnerability management, while Mandiant Threat Intelligence AI uses Sec-PaLM to rapidly detect and mitigate new threats.
- Bridging the Talent Gap: The Workbench helps non-experts understand security. While the Security Command Center AI offers clear attack graph visualizations and practical recommendations for risk mitigation, tools like Chronicle AI make it simple for users to discover and evaluate security incidents.

e. Microsoft Security Copilot [188, 199]

It is an AI-driven assistant designed to aid cybersecurity professionals in managing vast amounts of data and identifying security breaches. It integrates information from trusted sources such as the Cybersecurity and Infrastructure Security Agency, the National Institute of Standards and Technology's vulnerability database, and Microsoft's own threat intelligence network. Powered by OpenAI's GPT-4 and Microsoft's specialized security models, Copilot helps analysts with tasks like security investigations, summarizing events, analyzing files, URLs, code snippets, and incident information from various security tools. Key features of Microsoft Security Copilot include:

- User-Friendly Interface: The Copilot provides a simple prompt-based interface that allows security professionals to quickly gain insights or support for investigations and reports.
- Advanced Threat Intelligence: Leveraging 65 trillion daily data signals, Copilot assists in efficiently detecting and addressing threats by using Microsoft's robust threat intelligence system.
- Collaborative Workspace: Security teams can pin their findings into a shared workspace, facilitating collaboration and joint efforts in investigating and analyzing security issues.
- Prompt Book: This feature enables users to group tasks or automations into a single prompt, which simplifies complex processes like reverse-engineering scripts without needing an expert's involvement.
- Automated Reporting: Copilot can generate PowerPoint presentations that illustrate incidents and attack vectors, streamlining reporting processes.
- Feedback System: Users can provide feedback on incorrect results, helping to refine the system's accuracy and reduce errors over time.

IX. Metrics for Cyber security

Metrics for cyber security are classified into two classes namely, security operations centers (SOCs) metrics and performance metrics to evaluate the performance of ML/ DL approaches for forecast and prediction of cyber-attacks.

A. SOC Metric

A Security Operations Center (SOC) is a centralized hub that encompasses people, processes, and technology focused on the continuous monitoring, detection, and response to cybersecurity threats and incidents within an organization. The SOC's main goal is to protect the confidentiality, integrity, and availability of the organization's data and systems. It also plays a crucial role in enhancing cyber situational awareness, ensuring compliance, and managing threats effectively [200]. SOCs are being implemented by government agencies, universities, and both commercial and private organizations to protect their networks. However, most research on SOCs has been heavily centered on technology, often overlooking the human factors, operational processes, and the challenges faced by SOC analysts [201]. Table XI describes each cybersecurity performance metric and emphasizes the major features connected with it.

Table XI
Overview of cybersecurity SOCs metric

SL. no	Performance Metrics	Features	References
1	Number of security incidents	- Total number of incidents detected - Types of incidents (e.g., phishing, malware)	[202-206, 207-210,211-214, 217]
2	Mean time to reaction	- Average response time to incidents - Measures response efficiency	[207-209, 213-217]
3	Number of vulnerabilities	- Number of known vulnerabilities - Types of vulnerabilities (e.g., software, network)	[202-206, 209,216-217]
4	False-positive rate	- Proportion of non-threats identified as threats - Impact on resource allocation	[205- 207, 210-211, 213-214, 217]
5	Mean time to detect	- Average time to identify an incident - Indicates detection speed and capability	[205, 208-209, 213-214, 216-217]
6	Mean time to resolution	- Average time to resolve incidents - Measures incident management effectiveness	[204,207, 209, 212-213, 216-217]
7	Detections per category	- Number of detections by type (e.g., malware, phishing) - Helps in understanding threat distribution	[203, 206, 213-214, 217]
8	Mean time to vulnerability remedy	- Average time to patch or fix vulnerabilities - Reflects patch management process efficiency	[205-207, 209, 215, 217]
9	Number of vulnerable devices	- Total count of devices with vulnerabilities - Indicates network exposure and potential risk	[204-206, 217]

10	Incident avoidability	- Percentage of incidents that could have been prevented - Reflects proactive security measures	[206, 212, 216-217]
11	Number of monitored assets	- Total number of assets being monitored - Indicates scope of security operations	[204-205, 217]
12	Number of patched vulnerabilities	- Total number of vulnerabilities patched - Reflects patch management effectiveness	[203-204, 217]
13	Number of risks per severity	- Categorization of risks by severity level (e.g., high, medium, low) - Helps prioritize risk management	[204-205, 217]
14	Severity of security incidents	- Levels of severity assigned to incidents (e.g., critical, high, medium, low) - Affects response urgency	[205, 209, 213, 217]
15	Threat actor attribution	- Identification of sources behind attacks - Helps in understanding and mitigating specific threats	[205, 212,216-217]
16	Number of automated incidents	- Total incidents handled automatically by systems - Indicates the level of automation in incident response	[214-215, 217]
17	Mean time to escalation	- Average time taken to escalate an incident to higher-level support - Reflects incident management process	[209, 216-217]
18	Quality of eradication	- Effectiveness of measures taken to eliminate threats - Ensures threats are fully removed from the system	[212, 216-217]

B. Performance Metrics

Correctly evaluating the effectiveness of ML/DL models depends on the accurate interpretation of performance measures. The influence of the model may be ascertained in large part thanks to these measurements. AI models created for attack detection require careful consideration of several metrics [218-219]. The several measures used to evaluate ML/DL models within the parameters of the research examined are explained.

- True Positive (TP): The number of attacks that are correctly identified
- False Positive (FP): The number of benign instances mistakenly classified as attacks.
- True Negative (TN): The number of benign instances correctly identified.
- False Negative (FN): The number of attacks mistakenly classified as benign.

Table XII depicts the performance metrics derived from TP, FP, TN, and FN [218-219].

Table XII
Overview of performance metrics for assessing ML/ DL models for cyber-attack prediction

Performance metric	Explanation	Equation	Remarks
Recall (Sensitivity, Detection Rate)	To measure the effectiveness of a machine learning model in identifying all relevant instances of a specific class, particularly the positive class. In the context of attack detection, recall represents the proportion of actual attacks that the model successfully detects.	$Recall = \frac{TP}{TP + FN}$	A greater recall value means that the model has a lower rate of missed assaults (false negatives), indicating that it is effective in identifying the majority of actual attacks. Recall must be balanced with other metrics like precision because a high recall model may also result in more false positives.
Precision (True Positive Rate—TPR)	The proportion of predicted positives that are positive.	$Precision = \frac{TP}{TP + FP}$	Indicates the accuracy of positive predictions. High precision means that when the model predicts a positive class, it's likely to be correct.
Accuracy	The overall proportion of correctly classified instances (both positive and negative).	$Accuracy = \frac{TP + TN}{Total\ number\ of\ Instances}$	A broad measure of performance, but it can be misleading in imbalanced datasets.
False Alarm Rate (FAR) / False Positive Rate (FPR)	The proportion of negative instances that are incorrectly classified as positive.	$FPR = \frac{TP}{FP + TN}$	Indicates the likelihood of the model raising a false alarm, critical in systems where false positives are costly.
False Discovery Rate (FDR)	The proportion of positive predictions are incorrect.	$FDR = \frac{FP}{FP + TP}$	A lower FDR means that the majority of positive predictions are correct,

			increasing the model's reliability.
True Negative Rate (TNR)/ Specificity	The proportion of actual negatives that are correctly identified.	$TNR = \frac{TN}{TN+TP}$	Measures the model's ability to identify negatives correctly, crucial for minimizing false positives.
Negative Prediction Rate (NPR)	The proportion of predicted negatives are actually negative.	$NPR = \frac{TN}{TN+FN}$	Indicates the accuracy of negative predictions, ensuring that negative predictions are reliable.
Miss Rate—MR (False Negative Rate—FNR)	The proportion of positives that are incorrectly classified as negatives.	$MR = \frac{FN}{FN+TP}$	Measures the rate at which actual positives are missed by the model. A low miss rate is essential for high recall.
Error Rate (ER)	The overall proportion of incorrect predictions (both false positives and false negatives).	$ER = \frac{FP+FN}{TP+TN+FP+FN}$	Provides a general measure of the model's accuracy, reflecting how often it makes mistakes.
F- measure (F1-Score)	The harmonic mean of precision and recall, balancing the two metrics.	$F1-score = 2 * \frac{precision * recall}{precision + recall}$	Useful when you need a balance between precision and recall, especially in scenarios with an uneven class distribution. A high F1-Score indicates that the model has both high precision and recall.
Informedness [220-221]	It measures the probability of an informed decision.	TPR+TNR-1	A higher value signifies the model's strong ability to correctly identify positive instances (true positives) while avoiding false negatives. It reflects the model's capability to detect meaningful changes or anomalies in the dataset effectively.
Markedness[220-221]	It measures the consistency of predictions.	$precision + \frac{TN}{TN+FN} - 1$	A markedness value nearing the upper limit of its range indicates a highly reliable anomaly detection (AD) framework that effectively reduces false alerts.
MCC[220-221]	A balance measure of binary (normal or anomaly) classification quality.	$\frac{TP\ X\ TN - FP\ X\ FN}{\sqrt{(TP+FP)X\ (TP+FN)X\ (TN+FP)\ X\ (TN+FN)}}$	A higher MCC value represents a balanced and reliable model that performs well across all classes, even when the dataset is imbalanced

Informedness, Markedness, and Matthews Correlation Coefficient (MCC) are advanced performance metrics employed to evaluate ML, DL and LLM models. Markedness serves as a key measure of a model's ability to handle false positives (e.g., false alarms) and false negatives (e.g., missed detections). A high Markedness score indicates a reliable anomaly detection system capable of minimizing incorrect alerts, which is critical for ensuring stable and efficient substation operations by avoiding unnecessary interruptions. Informedness assesses the model's capacity to detect changes in dataset patterns that signal anomalies. For anomaly detection involving LLMs, where actual anomalies occur infrequently, MCC is particularly valuable. It provides an unbiased evaluation of the model's performance, ensuring it is not disproportionately affected by the majority class, thereby offering a clear and accurate measure of its overall effectiveness [220-221].

X. Observation and Challenges

In this work, we examined the state-of-the-art ML, DL, XAI, and GenAI techniques deployed in defending against various cyber-attacks and safeguarding diverse industrial cybersecurity domains. Although ML, DL, and XAI hold significant potential in fortifying cybersecurity domains, they encounter notable challenges in implementation. In the subsequent section, we discuss these challenges.

A. Datasets

An overview of prevalent and widely utilized datasets in ML and DL for various cyber-attacks and industries is presented

in Tables 1–3, respectively. However, a critical issue persists with many of these datasets: they lack updates in specific areas. This limitation may stem from concerns regarding privacy and ethics. Consequently, the most recent categories of cyber-attacks are often absent from publicly available datasets, hindering the efficacy of ML and DL models in training XAI applications to establish robust cyber-attack defense mechanisms. The ML, DL, XAI, and GENai models require large volumes of high-quality data for effective training. However, obtaining such data, particularly labeled datasets containing examples of cyber-attacks, can be challenging owing to privacy concerns and the scarcity of publicly available data.

B. Adversarial Attacks

The threat of adversarial attacks, in which malicious actors manipulate data to deceive ML and DL models, presents a significant challenge. Such attacks exploit vulnerabilities within models, resulting in misclassifications and undermining the efficacy of cyber-attack-prediction systems. For instance, adversaries may circumvent authentication systems such as XAI-enabled facial recognition systems or execute poisoning attacks to manipulate or corrupt training data sources [222]. To address these threats, a potential solution involves analyzing the "Desiderata for adversarial attacks in different scenarios involving explainable ML models," as outlined in [223].

One approach to mitigating adversarial attacks involves training the model to identify the inputs manipulated in such methods and responding to rejection. The model can learn to recognize input segments prone to carrying malicious information and evaluate potential consequences before withholding a response to a suspicious prompt. By training models to defend against adversarial attacks, we can instill trust in LLMs, ensuring that they do not inadvertently facilitate cybercriminals in obtaining malicious code.

C. Interpretability and Explainability

The XAI and GenAI techniques aim to provide insights into model predictions, but achieving interpretability in complex ML and DL models remains a challenge. Understanding the rationale behind model decisions is crucial to trust and accountability in cybersecurity applications.

D. Privacy and Ethical Issues

When integrating ML, DL, XAI, and GENAI within cybersecurity, it is crucial to address privacy and ethical concerns alongside technical challenges. Throughout the system life cycle, explicitly prioritizing privacy considerations is essential. Protecting individuals' privacy rights, particularly in sensitive areas, such as authentication and emails, is paramount. Moreover, these AI systems must undergo ethical scrutiny to mitigate biases and discrimination such as racism and sexism. Measures should ensure fairness in the decisions made and the explanations provided by AI systems. Eliminating ethical bias, particularly in specific cybersecurity domains, is imperative. Given that data originates from security-related sources, heightened privacy and security concerns necessitate safeguarding data and models from adversarial attacks and unauthorized access, ensuring that only authorized individuals have access to ML, DL, XAI, and GENAI models.

Ethical challenges in cybersecurity are crucial because they require maintaining the confidentiality of systems and information with respect to security, rights, and equality. Addressing these issues of ethics demands a diverse approach that involves ethical frameworks, behavior change interventions, and educational strategies.

Authors of [224] present several ethical frameworks for analyzing cybersecurity questions, emphasizing the importance of considering risk and probability. Ethical frameworks are crucial for addressing cybersecurity challenges. The principles and rights-based approaches, while valuable, require consideration of risk and probability [224]. Ethical impact assessments can help researchers evaluate their work's ethical implications [225]. However, current governance in cybersecurity ethics has shortfalls, particularly in the corporate sector, where research ethics boards are often unavailable [226]. To address these issues, ethics education in computer science curricula should be expanded, and effective codes of conduct should be developed [226]. The "ethics-by-design" approach in cybersecurity research emphasizes educating participants about ethical principles, discussing frameworks across stakeholders, and exploring techniques to apply ethical principles in research methodologies [225]. These efforts aim to improve ethical decision-making in both research and practice, addressing the complex ethical challenges posed by modern information and communication technologies.

Behavior change interventions are crucial for enhancing cybersecurity, as human behavior is often the weakest link in security protocols. Paper [227] highlights the need for ethical considerations in behavior change interventions aimed at improving cybersecurity, drawing from utilitarian, deontological, and virtue ethics traditions. While many organizations implement security awareness programs, these do not always lead to actual behavior change, highlighting the need for innovative techniques beyond mere awareness [228]. In healthcare, where cybersecurity risks can have dire consequences, structured behavior change techniques are essential to mitigate vulnerabilities among staff [229]. The AIDE approach—Assess, Identify, Develop, and Evaluate—provides a framework for implementing these interventions effectively. Additionally, understanding the factors influencing employee security behaviors is vital for designing successful interventions [230]. Overall, integrating ethical considerations and targeted behavior change strategies can significantly improve cybersecurity practices across various sectors.

Improving student engagement in professional ethics education, particularly in technical fields like cybersecurity, is crucial. Research suggests several effective strategies: aligning content with student interests, taking a pragmatic approach, addressing real-world complexities, and making content entertaining [231]. Authors in [232] proposed four principles for improving student engagement in professional ethics education, particularly cybersecurity, emphasizing real-world case studies and systemic perspectives. Classroom debates have been shown to stimulate affective learning, enhancing engagement, critical thinking, and ethical sensitivity [233]. The concept of 'practical wisdom' is proposed as an ethical framework for student engagement practices, with case studies highlighting various ethical challenges in research and teaching [234]. In ICT courses, where attrition rates are high, implementing a flipped-classroom approach and continuous assessment can increase student engagement in professional skills and ethics education, potentially improving academic performance and retention. These strategies collectively emphasize the importance of interactive, relevant, and ethically grounded approaches to teaching professional ethics in technical disciplines.

Authors of [235] discusses the evolving security and ethical challenges posed by information technology, noting the need for new laws and rules of acceptable conduct in the digital age. These approaches collectively aim to address the complex ethical landscape of cybersecurity, balancing individual, organizational, and societal interests.

The following are the key points to overcome the ethical issues in cybersecurity. The following are the key points to overcome the ethical issues in cybersecurity.

- Transparency: Ensure that security practices and policies are clear, accessible, and understandable to all stakeholders.
- Accountability: Establish clear lines of responsibility for cybersecurity actions and decisions.
- Data Privacy: Prioritize and protect the privacy of individuals' data by adhering to legal standards and best practices.
- Informed Consent: Obtain explicit consent from users before collecting, storing, or processing their data.
- Security Awareness Training: Educate employees and users on ethical practices, potential threats, and responsible behaviour in cybersecurity.
- Ethical Hacking: Use ethical hacking methods, such as penetration testing, to identify and address vulnerabilities before they can be exploited.
- Compliance with Regulations: Adhere to local, national, and international laws and regulations governing cybersecurity and data protection.
- Balancing Security and Freedom: Implement security measures that protect without unnecessarily infringing on personal freedoms or rights.
- Cultural Sensitivity: Consider cultural differences in the interpretation and application of cybersecurity practices.
- Continuous Monitoring and Improvement: Regularly review and update cybersecurity measures to ensure they remain ethical and effective in the face of evolving threats.

XI. Future Research Directions

The intersection of cybersecurity and AI is a dynamic and rapidly evolving field. While significant strides have been made in cyber-attack prediction using traditional ML, generative AI opens up new avenues for research and innovation. The same is represented in Figure 17. Below, we explain some promising future research directions:

Enhancing Generative AI for Cyber Attack Prediction

- **Hybrid Models**: Explore the fusion of generative and discriminative models for improved cyber-attack prediction accuracy. Generative models can create synthetic attack data to augment training datasets, improving the diversity and robustness of models; similarly, discriminative models excel at classification and prediction. By combining them, we can achieve more accurate and nuanced predictions. Potential architectures include generative adversarial networks (GANs) for data augmentation and support vector machines (SVMs) for classification [176, 236].
- **Adversarial Learning**: Develop adversarial techniques to strengthen the robustness of generative models against adversarial attacks, ensuring their reliability in real-world cyber security scenarios. We can improve their reliability in real-world scenarios by training models to defend against such attacks. Techniques include adversarial training, where models are exposed to adversarial examples during training [238].
- **Explainable AI (XAI)**: Explainable AI methods help understand the reasoning behind model predictions, which is crucial for identifying biases, debugging models, and gaining user acceptance. Techniques include Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). Explainable AI can focus on developing methods to interpret and explain the predictions made by generative AI models, building trust and transparency in the decision-making process.
- **Multimodal Learning**: Incorporate diverse data sources, such as network traffic, system logs, and threat intelligence feeds, into generative models to capture complex attack patterns and improve prediction accuracy. Incorporate network traffic, system logs, threat intelligence, and social media data. This holistic approach captures complex attack patterns and enhances model performance [237 – 238].

Advanced Threat Modeling and Simulation

- **Generative Threat Modelling**: Utilize generative AI to create realistic and diverse attack scenarios, enabling security teams to develop and test countermeasures proactively. Generative models can generate novel attack patterns, helping security teams anticipate threats, which helps develop and test countermeasures. Techniques include variational autoencoders (VAEs) and recurrent neural networks (RNNs) [239].
- **Red Team Automation**: Explore using generative AI to automate red team operations, generating novel attack techniques and strategies for assessing system vulnerabilities. Generative AI can assist red teams in developing innovative attack payloads, which helps organizations continuously evaluate their security posture. Challenges include ethical considerations and ensuring the generated attacks are realistic but not harmful [240 – 241].
- **Blue Team Optimisation**: Develop generative AI-driven tools to optimise defensive strategies based on simulated attack scenarios, improving security posture. Generative AI can generate different attack scenarios to test the effectiveness of defenses, which can lead to improved resource allocation and incident response plans. Techniques include reinforcement learning for decision-making [241 – 242].

Novel Data Sources and Feature Engineering

- **Unstructured Data Analysis**: Leverage generative AI to extract valuable information from unstructured data sources like social media, dark web forums, and code repositories to identify emerging threats [243 – 244].
- **Time Series Analysis**: Develop advanced time series analysis techniques using generative AI to capture long-term dependencies and trends in cyber-attack data. Generative models can help identify seasonal patterns, anomalies, and early warning signs. Techniques include long short-term memory (LSTM) networks and attention mechanisms [244 – 245].
- **Feature Learning**: Explore automated feature engineering methods using generative AI to discover hidden patterns and relationships within complex datasets. Generative models can learn meaningful data representations, reducing manual feature engineering efforts. Techniques include autoencoders and deep belief networks [245 – 246].

Ethical Considerations and Responsible AI

- **Privacy and Security**: Investigate privacy-preserving techniques for handling sensitive cyber security data while training generative AI models. Develop privacy-preserving techniques like differential privacy and federated learning. Address potential biases in data and models to ensure fairness [247 – 248].
- **Bias Mitigation**: Develop methods to address potential biases in generative AI models, ensuring fairness and equity in cyber-attack prediction [249].
- **Human-in-the-Loop**: Explore human-AI collaboration frameworks for cyber security, leveraging human expertise to guide and refine AI-generated predictions. Develop interactive systems where humans can provide feedback and refine model outputs [249-250]. Current approaches for increasing cybersecurity situational awareness (SA) either emphasize human expertise via alert configurations or instead consider only ML algorithms without human contributions. Yet, both methods have their weaknesses: human-based systems cannot prioritize which suspicious attempt is the more meaningful, while ML-based found alerts may be wrong, leading to decreased accuracy [250].

  One such human-in-the-loop active learning-based framework is proposed by researchers in [250], where they prioritize alerts based on their significance and utilize human investigation results to dynamically update an ever-improving detection system. These elements have been condensed into a framework of dynamic alert prioritization, human alert investigation, and incremental hypothesis testing. Analysts follow up on alerts generated by a Hidden Markov Model (HMM), and their feedback is leveraged to update the system's belief about the state of the attacker. The SA process is augmented with the manual expertise of a human based on policy, while a machine provides decision support, helping prioritize alerts and enhancing the accuracy of attack detection.

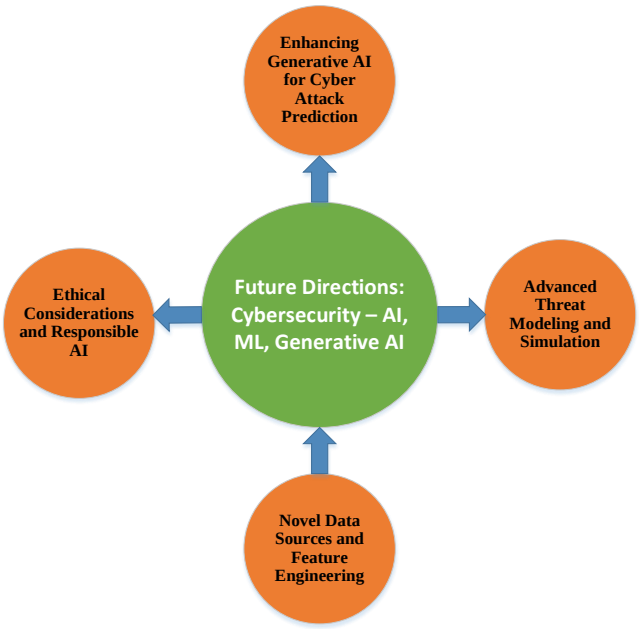

Figure 17. Future Directions: Cyberseurity – AI, ML and Generative AI.

XII. Conclusion

In cybersecurity, AI plays a pivotal role in analyzing datasets and monitoring diverse security threats and malicious activities. Effectively addressing myriad cybersecurity challenges, especially with the rising frequency of attacks, necessitates the integration of human expertise with AI capabilities. This study presents state-of-the-art benchmark cyber-attack datasets, ML and DL, and techniques for various cyber-attack predictions.

In the domain of cybersecurity, transparency, and explainability are essential for combating cyber threats and effectively analyzing security decisions. Hence, this study provides a thorough overview of cutting-edge research on XAI for cybersecurity applications. We delineate the fundamental principles and taxonomies of state-of-the-art XAI models, along with indispensable tools, such as a comprehensive framework and accessible datasets. We believe that this paper will be a valuable resource for researchers, developers, and security professionals seeking to leverage ML, DL, XAI, and GenAI models to address complex challenges within cybersecurity domains.

CONFLICTS OF INTEREST

‘‘The authors have no conflicts of interest to declare relevant to the content of this article”.

REFERENCES

[1] G Emile S, Mbungu Kala, “Critical Role of Cyber Security in Global Economy”, Open Journal of Safety Science and Technology, Vol. 13, pp. 231-248, 2023. doi: 10.4236/ojsst.2023.134012.

[2] Von Solms, Rossouw, and Johan Van Niekerk. "From information security to cyber security." computers & security, Vol. 38, pp. 97-102, 2013.

[3] J W Goodell and S. Corbet, Commodity market exposure to energy firm distress: Evidence from the colonial pipeline ransomware attack,’’Finance Res. Lett., vol. 51, Jan. 2023, Art. no. 103329

[4] R. Alkhadra, J. Abuzaid, M. AlShammari, and N. Mohammad, ‘‘Solar winds hack: In-depth analysis and countermeasures,’’ in Proc. 12th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT), Jul. 2021, pp. 1–7.

[5] Cobalt https://www.cobalt.io/blog/biggest-cybersecurity-attacks-in-history

[6] D.-Y. Kao, S.-C. Hsiao, and R. Tso, ‘‘Analyzing WannaCry ransomware considering the weapons and exploits,’’ in Proc. 21st Int. Conf. Adv. Commun. Technol. (ICACT), Feb. 2019, pp. 1098–1107.

[7] K. Bresniker, A. Gavrilovska, J. Holt, D. Milojicic and T. Tran, "Grand Challenge: Applying Artificial Intelligence and Machine Learning to Cybersecurity," in Computer, vol. 52, no. 12, pp. 45-52, Dec. 2019, doi: 10.1109/MC.2019.2942584.

[8] Husák, Martin, Jana Komárková, Elias Bou-Harb, and Pavel Čeleda. "Survey of attack projection, prediction, and forecasting in cyber security." IEEE Communications Surveys & Tutorials 21, no. 1 , 2018, pp. 640-660.

[9] Nachaat Mohamed, “Current trends in AI and ML for cybersecurity: A state-of-the-art survey”, Cogent Engineering, Vol. 10z, no. 2, 2023, DOI: 10.1080/23311916.2023.2272358

[10] L. Chan, I. Morgan, H. Simon, F. Alshabanat, D. Ober, J. Gentry, D. Min, and R. Cao, “Survey of ai in cybersecurity for information technology management,” in 2019 IEEE technology & engineering management conference (TEMSCON). IEEE, 2019, pp. 1–8.

[11] G. Disterer, “Iso/iec 27000, 27001 and 27002 for information security management,” 2013.

[12] Hua Li J. Cyber security meets artificial intelligence: a survey. Front Inf Technol Electron Eng. 2018;19(12):1462–74. https://doi.org/10.1631/FITEE.1800573.

[13] Sarker IH, Kayes ASM, Badsha S, Alqahtani H, Watters P, Ng A. Cybersecurity data science: an overview from machine learning perspective. J Big Data. 2020. https://doi.org/10.1186/s40537-020-00318-5.

[14] Aggarwal, Deepshikha, Deepti Sharma, and Archana B. Saxena. "Role of AI in cyber security through Anomaly detection and Predictive analysis." *Journal of Informatics Education and Research* 3, no. 2 (2023).

[15] Srivastava, Gautam, Rutvij H. Jhaveri, Sweta Bhattacharya, Sharnil Pandya, Praveen Kumar Reddy Maddikunta, Gokul Yenduri, Jon G. Hall, Mamoun Alazab, and Thippa Reddy Gadekallu. "XAI for cybersecurity: state of the art, challenges, open issues and future directions." *arXiv preprint arXiv:2206.03585* (2022).

[16] I. H. Sarker, M. H. Furhad and R. Nowrozy, "AI-driven cybersecurity: An overview security intelligence modeling and research directions", *Social Netw. Comput. Sci.*, vol. 2, no. 3, pp. 1-18, May 2021.

[17] D. Ucci, L. Aniello and R. Baldoni, "Survey of machine learning techniques for malware analysis", *Comput. Secur.*, vol. 81, pp. 123-147, Mar. 2019.

[18] D. Kwon, H. Kim, J. Kim, S. C. Suh, I. Kim and K. J. Kim, "A survey of deep learning-based network anomaly detection", *Cluster Comput.*, vol. 22, pp. 949-961, Jan. 2019.

[19] R. A. Nafea and M. A. Almaiah, "Cyber security threats in cloud: Literature review", *Proc. Int. Conf. Inf. Technol. (ICIT)*, pp. 779-786, Jul. 2021.

[20] A. Kuppa and N.-A. Le-Khac, "Black box attacks on explainable artificial Intelligence(XAI) methods in cyber security", *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, pp. 1-8, Jul. 2020.

[21] K. D. Ahmed and S. Askar, "Deep learning models for cyber security in IoT networks: A review", *Int. J. Sci. Bus.*, vol. 5, no. 3, pp. 61-70, 2021.

[22] J. Gerlings, A. Shollo and I. Constantiou, "Reviewing the need for explainable artificial intelligence (xAI)", *arXiv:2012.01007*, 2020.

[23] G. Jaswal, V. Kanhangad and R. Ramachandra, AI and Deep Learning in Biometric Security: Trends Potential and Challenges, Boca Raton, FL, USA:CRC Press, 2021.

[24] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", *arXiv:1811.10154*, 2018.

[25] Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," in *IEEE Access*, vol. 10, pp. 93104-93139, 2022, doi: 10.1109/ACCESS.2022.3204051

[26] A. Bandi, P. V. S. R. Adapa, and Y. E. V. P. K. Kuchi, “The power of generative ai: A review of requirements, models, input-output formats, evaluation metrics, and challenges,” Future Internet, vol. 15, no. 8, 2023. [Online]. Available: https://www.mdpi.com/1999-5903/15/8/260

[27] J. Babcock and R. Bali, Generative AI with python and tensorflow 2: Create images, text, and music with Vaes, Gans, LSTMs, Transformer models. Packt Publishing, Limited, 2021

[28] “Chatgpt,” https://chat.openai.com/, (Accessed on 06/22/2023). [24]

[29] “Dall·e now available without waitlist,” https://openai.com/blog/dall-e-now-available-without-waitlist, (Accessed on 06/22/2023).

[30] [25] B. Hitaj, P. Gasti, G. Ateniese, and F. Perez-Cruz, “Passgan: A deep learning approach for password guessing,” 2019.

[31] P. Dhoni and R. Kumar, “Synergizing generative ai and cybersecurity: Roles of generative ai entities, companies, agencies, and government in enhancing cybersecurity,” Aug. 2023. [Online]. Available: http://dx.doi.org/10.36227/techrxiv.23968809.v1

[32] Plotnek, Jordan J., and Jill Slay. "Cyber terrorism: A homogenized taxonomy and definition." Computers & Security, Vol. 102, 2021, 102145

[33] Kim, Seungmin, Gyunyoung Heo, Enrico Zio, Jinsoo Shin, and Jae-gu Song. "Cyber attack taxonomy for digital environment in nuclear power plants." Nuclear Engineering and Technology 52, no. 5,2020, pp. 995-1001.

[34] Wu, Mingtao, and Young B. Moon. "Taxonomy of cross-domain attacks on cybermanufacturing system." Procedia Computer Science, no. 114, 2017, pp. 367-374.

[35] Heartfield, Ryan, George Loukas, Sanja Budimir, Anatolij Bezemskij, Johnny RJ Fontaine, Avgoustinos Filippoupolitis, and Etienne Roesch, "A taxonomy of cyber-physical threats and impact in the smart home", Computers & Security, Vol. 78, 2018, pp. 398-428.

[36] Alkhalil Z, Hewage C, Nawaf L and Khan I, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy", Front. Comput. Sci. Vol. 3, no. 563060, 2021, pp. 1-23.

[37] N. Z. Gorment, A. Selamat, L. K. Cheng and O. Krejcar, "Machine Learning Algorithm for Malware Detection: Taxonomy, Current Challenges, and Future Directions," in IEEE Access, vol. 11, pp. 141045-141089, 2023, doi: 10.1109/ACCESS.2023.3256979.

[38] Ferrag, Mohamed Amine, Leandros Maglaras, Sotiris Moschoyiannis, and Helge Janicke. "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study." Journal of Information Security and Applications, Vol. 50, no. 102419, 2021, pp. 2020.

[39] Markus Ring, Sarah Wunderlich, Deniz Scheuring, Dieter Landes, Andreas Hotho, "A survey of network-based intrusion detection data sets," Computers & Security, Vol. 86, 2019, pp. 147-167,

[40] 1998 DARPA intrusion detection evaluation dataset | MIT lincoln laboratory, 2020, https://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusion-detection-evaluation-dataset

[41] KDD cup 1999 data, The UCI KDD Archive, Information and Computer Science, University of California, Irvine, Irvine, CA 92697-3425 Last modified: October 28, 1999, URL: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.htm

[42] Sperotto A, Sadre R, Van Vliet F, Pras A., "A labeled data set for flow-based intrusion detection", In: Proceedings of the international workshop on IP operations and management. Springer; 2009. p. 39–50. doi:10.1007/978-3-642-04968-2_4.

[43] Benjamin Sangster, T. J. O'Connor, Thomas Cook, Robert Fanelli, Erik Dean, William J. Adams, Chris Morrell, and Gregory Conti., "Toward instrumenting network warfare competitions to generate labeled datasets", . In Proceedings of the 2nd conference on Cyber security experimentation and test (CSET'09). USENIX Association, USA, 2009.

[44] Gringoli F, Salgarelli L, Dusi M, Cascarano N, Risso F, et al., " GT: picking up the truth from the ground for internet traffic", ACM SIGCOMM Comput Commun Rev, Vol. 39, no. 5, 2009, pp. 12–18. doi:10.1145/1629607.1629610.

[45] Saad S, Traore I, Ghorbani A, Sayed B, Zhao D, Lu W, Felix J, Hakimian P, "Detecting P2P botnets through network behavior analysis and machine learning", In: Proceedings of the international conference on privacy, security and trust (PST). IEEE, 2011. pp. 174–80. doi:10.1109/PST.2011.5971980

[46] Bhattacharya S, Selvakumar S, "SSENet-2014 Dataset: A Dataset for Detection of Multiconnection Attacks. In: Proceedings of the international conference on eco-friendly computing and communication systems (ICECCS)", IEEE, 2014. pp. 121–126. doi:10.1109/Eco-friendly.2014.100

[47] Jazi HH, Gonzalez H, Stakhanova N, Ghorbani AA, "Detecting HTTP-based application layer DoS attacks on web servers in the presence of sampling", Comput Netw, Vol. 121, 2017, pp. 25–36. doi:10.1016/j.comnet.2017.03.018

[48] Shiravi A, Shiravi H, Tavallaee M, Ghorbani AA, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection", Comput Secur, Vol. 31, no. 3, 2012, pp. 357–374. doi:10.1016/j.cose.2011.12.012.

[49] Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K, "Towards Generating Real-life Datasets for Network Intrusion Detection", Int. J. Netw. Secur., Vol. 17, no. 6, 2015, pp. 683-701.

[50] Santanna JJ, van Rijswijk-Deij R, Hofstede R, Sperotto A, Wierbosch M, Granville LZ, Pras A, "Booters - An analysis of DDoS-as-a-service attacks", In: Proceedings of the IFIP/IEEE international symposium on integrated network management (IM); 2015. Pp. 243–251. doi:10.1109/INM.2015.7140298

[51] Garcia S, Grill M, Stiborek J, Zunino A, "An empirical comparison of botnet detection methods", Comput Secur, Vol.45, 2014, pp. 100–123. doi:10.1016/j.cose.2014.05.011.

[52] Hofstede R, Hendriks L, Sperotto A, Pras A, "SSH compromise detection using NetFlow/IPFIX. ACM SIGCOMM", Comput Commun Rev, vol. 44, no. 5, 2014, pp. 20–26. doi:10.1145/2677046.2677050

[53] Beigi EB, Jazi HH, Stakhanova N, Ghorbani AA, "Towards effective feature selection in machine learning-based botnet detection approaches. In: Proceedings of the IEEE conference on communications and network security. IEEE; 2014. p. 247–55. doi:10.1109/CNS.2014.6997492

[54] Wheelus C, Khoshgoftaar TM, Zuech R, Najafabadi MM, "A session based approach for aggregating network traffic data - The SANTA dataset", In: Proceedings of the IEEE international conference on bioinformatics and bioengineering (BIBE). IEEE, 2014. pp. 369–378. doi:10.1109/BIBE.2014.72

[55] Kolias C, Kambourakis G, Stavrou A, Gritzalis S, " Intrusion detection in 802.11 networks: empirical evaluation of threats and a public dataset", IEEE Commun Surv Tutor,Vol. 18, no. 1, 2016, pp. 184–208. doi:10.1109/COMST.2015.2402161.

[56] Zuech R, Khoshgoftaar TM, Seliya N, Najafabadi MM, Kemp C, "A new intrusion detection benchmarking system", In: Proceedings of the international florida artificial intelligence research society conference (FLAIRS). AAAI Press, 2015. p. 252–256.

[57] Moustafa N, Slay J, "UNSW-NB15: a comprehensive data set for network intrusion detection systems", In: Proceedings of the military communications and information systems conference (MilCIS), IEEE, 2015. pp. 1–6

[58] Alkasassbeh M, Al-Naymat G, Hassanat A, Almseidin M, "Detecting distributed denial of service attacks using data mining techniques", Int J Adv Comput Sci Appl (IJACSA), Vol. 7, no. 1, 2016, pp. 436–445.

[59] Beer F, Hofer T, Karimi D, Bühler U., "A new Attack Composition for Network Security", In: 10. DFN-Forum Kommunikationstechnologien. Gesellschaft für Informatik eV;, 2017. pp. 11–20

[60] Sharafaldin I, Lashkari AH, Ghorbani AA, " Toward generating a new intrusion detection dataset and intrusion traffic characterization", In: Proceedings of the international conference on information systems security and privacy (ICISSP), 2018, pp. 108–16. doi:10.5220/0006639801080116.

[61] Ring M, Landes D, Hotho A, "Detection of slow port scans in flow-based network traffic", PLOS ONE, Vol. 13, no. 9, 2018, 1–18. doi: 10.1371/journal.pone.0204507.

[62] Ring M, Wunderlich S, Grüdl D, Landes D, Hotho A, "Creation of flow-based data sets for intrusion detection", J Inf Warf, Vol. 16, 2017, pp. 40–53.

[63] Turcotte, Melissa JM, Alexander D. Kent, and Curtis Hash. "Unified host and network data set." In Data science for cyber-security, 2019, pp. 1-22.

[64] Maciá-Fernández G, Camacho J, Magán-Carrión R, García-Teodoro P, Therón R, "UGR'16: a new dataset for the evaluation of cyclostationarity-based network IDSs", Comput Secur, Vol. 73, 2018, pp. 411–24. doi:10.1016/j.cose.2017.11.004

[65] Damasevicius, Robertas, Algimantas Venckauskas, Sarunas Grigaliunas, Jevgenijus Toldinas, Nerijus Morkevicius, Tautvydas Aleliunas, and Paulius Smuikys, "LITNET-2020: An Annotated Real-World Network Flow Dataset for Network Intrusion Detection" Electronics, Vol. 9, no. 5: 800, 2020, pp. 1-23, https://doi.org/10.3390/electronics9050800

[66] Ferriyan, Andrey, Achmad Husni Thamrin, Keiji Takeda, and Jun Murai. "Generating Network Intrusion Detection Dataset Based on Real and Encrypted Synthetic Attack Traffic" Applied Sciences, Vol. 11, no. 17: 7868., 2021. https://doi.org/10.3390/app11177868

[67] Elif Değirmenci, Yunus Sabri Kırca, İlker Özçelik, Ahmet Yazıc, " ROSIDS23: Network intrusion detection dataset for robot operating system", Data in Brief, Vol. 51, 2023, pp. 1-12. https://doi.org/10.1016/j.dib.2023.109739.

[68] Mihailescu M-E, Mihai D, Carabas M, Komisarek M, Pawlicki M, Hołubowicz W, Kozik R, "The Proposition and Evaluation of the RoEduNet-SIMARGL2021 Network Intrusion Detection Dataset", Sensors., Vol. 21, no. 13:4319, 2021, https://doi.org/10.3390/s21134319

[69] Mohammad Almseidin, Jamil Al-Sawwa, Mouhammd Alkasassbeh, June 18, 2022, "Multi-Step Cyber-Attack Dataset (MSCAD for Intrusion Detection)", IEEE Dataport, doi: https://dx.doi.org/10.21227/phr0-e264.

[70] AF. Yazı, FÖ Çatak, E. Gül, Classification of Metamorphic Malware with Deep Learning (LSTM), IEEE Signal Processing and Applications Conference, 2019.

[71] D. S. Keyes, B. Li, G. Kaur, A. H. Lashkari, F. Gagnon and F. Massicotte, "EntropLyzer: Android Malware Classification and Characterization Using Entropy Analysis of Dynamic Characteristics," 2021 Reconciling Data Analytics, Automation, Privacy, and Security: A Big Data Challenge (RDAAPS), Hamilton, ON, Canada, 2021, pp. 1-12, doi: 10.1109/RDAAPS48126.2021.9452002.

[72] Arash Habibi Lashkari, Andi Fitriah A. Kadir, Laya Taheri, and Ali A. Ghorbani, "Toward Developing a Systematic Approach to Generate Benchmark Android Malware Datasets and Classification", In the proceedings of the 52nd IEEE International Carnahan Conference on Security Technology (ICCST), Montreal, Quebec, Canada, 2018.

[73] Arash Habibi Lashkari, Andi Fitriah A. Kadir, Hugo Gonzalez, Kenneth Fon Mbah and Ali A. Ghorbani, "Towards a Network-Based Framework for Android Malware Detection and Characterization", In the proceeding of the 15th International Conference on Privacy, Security and Trust, PST, Calgary, Canada, 2017

[74] Harang, Richard, and Ethan M. Rudd. "SOREL-20M: A large scale benchmark dataset for malicious PE detection." arXiv preprint arXiv:2012.07634 (2020).

[75] Alibaba Cloud Malware Detection Based on Behaviors. 2021. Available online: https://tianchi.aliyun.com/competition/entrance/231694/information (accessed on 20 June 2021).

[76] Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao, "A survey on security and privacy issues in Internet-of-Things," IEEE Internet Things J., vol. 4, no. 5, pp. 1250–1258, Oct. 2017.

[77] F. De Keersmaeker, Y. Cao, G. K. Ndonda and R. Sadre, "A Survey of Public IoT Datasets for Network Security Research," in IEEE Communications Surveys & Tutorials, vol. 25, no. 3, pp. 1808-1840, thirdquarter 2023, doi: 10.1109/COMST.2023.3288942.

[78] https://www.stratosphereips.org/datasets-iot23, (accessed on 20 June 2021).

[79] Alsaedi, Abdullah, Nour Moustafa, Zahir Tari, Abdun Mahmood, and Adnan Anwar. "TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems." Ieee Access 8 (2020): 165130-165150.

[80] Hindy, H., Bayne, E., Bures, M., Atkinson, R., Tachtatzis, C., Bellekens, X. "Machine Learning Based IoT Intrusion Detection System: An MQTT Case Study (MQTT-IoT-IDS2020 Dataset)", . In: Ghita, B., Shiaeles, S. (eds) Selected Papers from the 12th International Networking Conference. INC 2020. Lecture Notes in Networks and Systems, vol 180. Springer, Cham. https://doi.org/10.1007/978-3-030-64758-2_6

[81] Ferrag, Mohamed Amine, Othmane Friha, Djallel Hamouda, Leandros Maglaras, and Helge Janicke. "Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning." IEEE Access 10 (2022): 40281-40306.

[82] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, ''N-baiot—Network-based detection of IoT botnet attacks using deep autoencoders,'' IEEE Pervasive Comput., vol. 17, no. 3, pp. 12–22, Oct. 2018

[83] M. Zolanvari, M. A. Teixeira, L. Gupta, K. M. Khan, and R. Jain, ''Machine learning-based network vulnerability analysis of industrial Internet of Things,'' IEEE Internet Things J., vol. 6, no. 4, pp. 6822–6834, Aug. 2019.

[84] M. Al-Hawawreh, E. Sitnikova, and N. Aboutorab, ''X-IIoTID: A connectivity-agnostic and device-agnostic intrusion data set for industrial Internet of Things,'' IEEE Internet Things J., vol. 9, no. 5, pp. 3962–3977, Mar. 2022.

[85] Nickolaos Koroniotis, Nour Moustafa, Elena Sitnikova, Benjamin Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," Future Generation Computer Systems, Volume 100, 2019, PP. 779-796, ISSN 0167-739X, https://doi.org/10.1016/j.future.2019.05.041.

[86] Draper-Gil, Gerard, Arash Habibi Lashkari, Mohammad Saiful Islam Mamun, and Ali A. Ghorbani. "Characterization of encrypted and vpn traffic using time-related." In *Proceedings of the 2nd international conference on information systems security and privacy (ICISSP)*, pp. 407-414. 2016.

[87] Habibi Lashkari, Arash, Gurdip Kaur, and Abir Rahali. "Didarknet: A contemporary approach to detect and characterize the darknet traffic using deep image learning." In *Proceedings of the 2020 10th International Conference on Communication and Network Security*, pp. 1-13. 2020.

[88] Stewart, Emma, Anna Liao, and Ciaran Roberts. "Open μpmu: A real world reference distribution micro-phasor measurement unit data set for research and application development." (2016).

[89] Hines, Paul, Seth Blumsack, E. Cotilla Sanchez, and Clayton Barrows. "The topological and electrical structure of power grids." In *2010 43rd Hawaii International Conference on System Sciences*, pp. 1-10. IEEE, 2010.

[90] Adhikari U., Pan S., Morris T., Borges R., Beave J.. Industrial Control System (ICS) Cyber Attack Datasets. https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets. last accessed 25 November 2024.

[91] Umass Dataset. http://traces.cs.umass.edu. last accessed 25 November 2024.

[92] Lashkari, Arash Habibi, Gerard Draper Gil, Mohammad Saiful Islam Mamun, and Ali A. Ghorbani. "Characterization of tor traffic using time based features." In *International Conference on Information Systems Security and Privacy*, vol. 2, pp. 253-262. SciTePress, 2017.

[93] Fontugne, Romain, Pierre Borgnat, Patrice Abry, and Kensuke Fukuda. "Mawilab: combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking." In *Proceedings of the 6th International COnference*, pp. 1-12. 2010.

[94] M. Ozkan-Okay et al., "A Comprehensive Survey: Evaluating the Efficiency of Artificial Intelligence and Machine Learning Techniques on Cyber Security Solutions," in IEEE Access, vol. 12, pp. 12229-12256, 2024, doi: 10.1109/ACCESS.2024.3355547.

[95] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed and M. Xu, "A Survey on Machine Learning Techniques for Cyber Security in the Last Decade," in IEEE Access, vol. 8, pp. 222310-222354, 2020, doi: 10.1109/ACCESS.2020.3041951.

[96] M. F. Franco, E. Sula, A. Huertas, E. J. Scheid, L. Z. Granville and B. Stiller, "SecRiskAI: a Machine Learning-Based Approach for Cybersecurity Risk Prediction in Businesses," 2022 IEEE 24th Conference on Business Informatics (CBI), Amsterdam, Netherlands, 2022, pp. 1-10, doi: 10.1109/CBI54897.2022.00008.

[97] M. C. Belavagi and B. Muniyal, "Performance evaluation of supervised machine learning algorithms for intrusion detection,'' Proc. Comput. Sci., vol. 89, pp. 117–123, Jan. 2016.

[98] J. Camacho, G. Maciá-Fernández, N. M. Fuentes-García, and E. Saccenti, ''Semi-supervised multivariate statistical network monitoring for learning security threats,'' IEEE Trans. Inf. Forensics Security, vol. 14, no. 8, pp. 2179–2189, Aug. 2019.

[99] H. Singh, Performance analysis of unsupervised machine learning techniques for network traffic classification,'' in Proc. 5th Int. Conf. Adv. Comput. Commun. Technol., Feb. 2015, pp. 401–404.

[100]] T. T. Nguyen and V. J. Reddi, ''Deep reinforcement learning for cyber security,'' IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 8, pp. 1–17, Nov. 2021.

[101] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. Vol. 2, No. 160, 2021. https://doi.org/10.1007/s42979-021-00592-x

[102] Sarker, I.H., Furhad, M.H. & Nowrozy, R, "AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions", SN COMPUT. SCI. Vol. 2, No. 173, 2021. https://doi.org/10.1007/s42979-021-00557-0

[103] Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Fast algorithms for mining association rules. In: Proceedings of the International Joint Conference on Very Large Data Bases, Santiago Chile. 1994; 1215: 487–499.

[104] Han J, Pei J, Yin Y., "Mining frequent patterns without candidate generation", In: ACM Sigmod Record, ACM, Vol. 29, pp. 1–12, 2000.

[105] Liu H, Motoda H, "Feature extraction, construction and selection: A data mining perspective", Springer Science & Business Media, vol. 453, 1998.
[106] Puterman ML. Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons; 2014.
[107] Kaelbling LP, Littman ML, Moore AW, "Reinforcement learning: a survey", J Artif Intell Res. Vol. 4, pp. 237-28, 1996.
[108] Amjad, Naeem, Hammad Afzal, Muhammad Faisal Amjad, and Farrukh Aslam Khan. "A multi-classifier framework for open source malware forensics." In 2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), pp. 106-111. IEEE, 2018.
[109] Srinivasan, Sathiyandrakumar, and P. Deepalakshmi. "ENetRM: ElasticNet Regression Model based malicious cyber-attacks prediction in real-time server." Measurement: Sensors, 25, 2023, 100654.
[110] Chesney, Steve, Kaushik Roy, and Sajad Khorsandroo. "Machine learning algorithms for preventing IoT cybersecurity attacks." In Intelligent Systems and Applications: Proceedings of the 2020 Intelligent Systems Conference (IntelliSys) Volume 3, pp. 679-686. Springer International Publishing, 2021.
[111] Althagafi, Safaa Sharif, Hajer Fahad Aljudiaibi, Badriah Abdelhadi Alharbi, and Raniyah Wazirali. "Uses of Artificial Intelligence in Cyber Security to Mitigate DDOS." In Proceedings of the Future Technologies Conference, pp. 550-565. Cham: Springer Nature Switzerland, 2023.
[112] Ahmed, Yussuf, A. Taufiq Asyhari, and Md Arafatur Rahman. "A cyber kill chain approach for detecting advanced persistent threats", Computers, Materials and Continua, Vol. 67, no. 2, 2497-2513, 2021.
[113] Balakrishnan, Yeshaswini, and P. N. Renjith. "An analysis on Keylogger Attack and Detection based on Machine Learning", In 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), pp. 1-8. IEEE, 2023.
[114] Haji, Saad Hikmat, and Siddeeq Y. Ameen. "Attack and anomaly detection in iot networks using machine learning techniques: A review." Asian journal of research in computer science 9, no. 2 , PP. 30-46, 2021
[115] Binu, P. K., and M. Kiran. "Attack and Anomaly Prediction in IoT Networks using Machine Learning Approaches." In 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1-6. IEEE, 2021.
[116] Agarwal, Anshul. "Load forecast anomaly detection under cyber-attacks using a novel approach." In 2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), pp. 1-6. IEEE, 2022.
[117] Diaba, Sayawu Yakubu, Miadreza Shafie-Khah, and Mohammed Elmusrati. "Cyber Security in Power Systems Using Meta-Heuristic and Deep Learning Algorithms." IEEE Access, Vol. 11, 18660-18672, 2023
[118] Ünözkan, Hüseyin, Mehmet Ertem, and Salaheddine Bendak. "Using attack graphs to defend healthcare systems from cyberattacks: a longitudinal empirical study." Network Modeling Analysis in Health Informatics and Bioinformatics, Vol. 11, no. 1 (2022)
[119] Tomer, Vikas, and Sachin Sharma. "Detecting iot attacks using an ensemble machine learning model." Future Internet 14, no. 4, 2022.
[120] Alsamiri, Jadel, and Khalid Alsubhi. "Internet of things cyber attacks detection using machine learning." International Journal of Advanced Computer Science and Applications Vol. 10, no. 12, 2019.
[121] Dutta, Chiranjit, M. Maheswari, K. G. Saravanan, Navdeep Dhaliwal, Akhilesh Pandey, and S. Sophia. "Prediction and Analysis of Various Cyber Attack Models in Cyber Physical System in Virtual Environment." In 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), pp. 1260-1264. IEEE, 2023.
[122] Swaminathan, Aravind, Balamurali Ramakrishnan, M. Kanishka, and R. Surendran. "Prediction of Cyber-attacks and Criminality Using Machine Learning Algorithms." In 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), pp. 547-552. IEEE, 2022.
[123] Macas, Mayra, Chunming Wu, and Walter Fuertes. "A survey on deep learning for cybersecurity: Progress, challenges, and opportunities." Computer Networks 212 (2022): 109032.
[124] Apruzzese, Giovanni, Michele Colajanni, Luca Ferretti, Alessandro Guido, and Mirco Marchetti. "On the effectiveness of machine and deep learning for cyber security." In 2018 10th international conference on cyber Conflict (CyCon), pp. 371-390. IEEE, 2018.
[125]. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, "Densely connected convolutional networks", in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 4700–4708, http://dx.doi.org/10. 1109/cvpr.2017.243.
[126] R. Pascanu, C. Gulcehre, K. Cho, Y. Bengio, "How to construct deep recurrent neural networks,", 2013, arXiv preprint arXiv:1312.6026.
[127] D.P. Kingma, M. Welling, "Auto-encoding variational bayes", 2013, arXiv preprint arXiv:1312.6114.
[128] .G. E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput., Vol. 18, no, 7,pp. 1527–1554, 2006. http://dx.doi.org/10.1162/ neco.2006.18.7.1527.
[129] G.-J. Qi, Loss-sensitive generative adversarial networks on Lipschitz densities, Int. J. Comput. Vis. , Vol. 128, no. 5, pp. 1118-1140, 2019. http://dx.doi.org/10.1007/ s11263-019-01265-2.
[130] V. Mnih, A.P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, K. Kavukcuoglu, Asynchronous methods for deep reinforcement learning, in: International Conference on Machine Learning, 2016, pp. 1928–1937.
[131] Kravchik, Moshe, and Asaf Shabtai. "Detecting cyber attacks in industrial control systems using convolutional neural networks." In Proceedings of the 2018 workshop on cyber-physical systems security and privacy, pp. 72-83. 2018.
[132] Hussain, Bilal, Qinghe Du, Bo Sun, and Zhiqiang Han. "Deep learning-based DDoS-attack detection for cyber–physical system over 5G network." IEEE Transactions on Industrial Informatics 17, no. 2 (2020): 860-870.
[133] Balogun, Bukola & Tripathi, Khushboo & Tiwari, Shrikant & S, Shyam & Tyagi, Amit, " A blockchain-based deep learning approach for cyber security in next-generation medical cyber-physical systems", Journal of Autonomous Intelligence, Vol. 7. 1478, 2024 10.32629/jai.v7i5.1478.
[134] Luo, Yuan, Ya Xiao, Long Cheng, Guojun Peng, and Danfeng Yao. "Deep learning-based anomaly detection in cyber-physical systems: Progress and opportunities." ACM Computing Surveys (CSUR), Vol. 54, no. 5, 2021, pp. 1-36.
[135] Dalal, S., Manoharan, P., Lilhore, U.K. et al, "Extremely boosted neural network for more accurate multi-stage Cyber-attack prediction in cloud computing environment", Journal of Cloud Computing, Vol. 12, no. 14 , 2023. https://doi.org/10.1186/s13677-022-00356-9
[136] S. Mahdavifar and A. A. Ghorbani, Dennes: Deep embedded neural network expert system for detecting cyber-attacks,'' Neural Comput. 2200 Appl., vol. 32, no. 18, pp. 14753–14780, 2020
[137] Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PloS one 10, no. 7 (2015): e0130140.
[138] Parsa A.B., Movahedi A., Taghipour H., Derrible S., Mohammadian A. (Kouros) "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis", Accident Analysis & Prevention, Vol. 136: 105405, 2020. doi: 10.1016/j.aap.2019.105405.
[139] P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, and M. A. Przybocki, ''Four principles of explainable artificial intelligence,'' 1643 NIST Interagency, Gaithersburg, MD, USA, Internal Rep. NISTIR-8312, 1644 Aug. 2020, doi: 10.6028/NIST.IR.8312.
[140] Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," in IEEE Access, vol. 10, pp. 93104-93139, 2022, doi: 10.1109/ACCESS.2022.3204051.
[141] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ''Grad-CAM: Visual explanations from deep networks via gradient-based localization,'' Int. J. Comput. Vis., vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7

[142] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting predictions," in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 1–10.
[143] R. Iyer, Y. Li, H. Li, M. Lewis, R. Sundar, and K. Sycara, "Transparency and explanation in deep reinforcement learning neural networks,'' 2018, arXiv:1809.06061.
[144] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. Vera Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, ''One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques,'' 2019, arXiv:1909.03012.
[145] S. Bose, T. Barao, and X. Liu, ''Explaining AI for malware detection: Analysis of mechanisms of MalConv,'' in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2020, pp. 1–8.
[146] E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, and C. K. Nicholas. (Jun. 2018). Malware Detection by Eating a Whole EXE. Accessed: Jul. 18, 2022. [Online]. Available: https://www.aaai.org/ocs/index.php/WS/AAAIW18/paper/view/16422
[147] H. S. Anderson and P. Roth, EMBER: An open dataset for training static PE malware machine learning models,'' 2018, arXiv:1804.04637.
[148] B. Wu, S. Chen, C. Gao, L. Fan, Y. Liu, W. Wen, and M. R. Lyu, ''Why an Android APP is classified as malware: Toward malware classification interpretation,'' ACM Trans. Softw. Eng. Methodol., vol. 30, no. 2, pp. 1–29, Apr. 2021.
[149] W. Han, J. Xue, Y. Wang, L. Huang, Z. Kong, and L. Mao, ''MalDAE: Detecting and explaining malware based on correlation and fusion of static and dynamic characteristics,'' Comput. Secur., vol. 83, pp. 208–233, Jun. 2019.
[150] H. Suryotrisongko, Y. Musashi, A. Tsuneda, and K. Sugitani, ''Robust botnet DGA detection: Blending XAI and OSINT for cyber threat intel- 2279 ligence sharing,'' IEEE Access, vol. 10, pp. 34613–34624, 2022.
[151] X. Zhu, Y. Zhang, Z. Zhang, D. Guo, Q. Li, and Z. Li, ''Interpretability evaluation of botnet detection model based on graph neural network,'' in Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS), 2292 May 2022, pp. 1–6
[152] M. Kouvela, I. Dimitriadis, and A. Vakali, ''Bot-detective: An explainable Twitter bot detection service with crowdsourcing functionalities,'' in Proc. 12th Int. Conf. Manage. Digit. EcoSystems, Nov. 2020, pp. 55–63. 2314
[153] C. Khanan, W. Luewichana, K. Pruktharathikoon, J. Jiarpakdee, C. Tantithamthavorn, M. Choetkiertikul, C. Ragkhitwetsagul, and T. Sunetnanta, ''JITBOT: An explainable just-in-time defect prediction bot,'' in Proc. 35th IEEE/ACM Int. Conf. Automated Softw. Eng., Sep. 2020, pp. 1336–1339.
[154] I. Dimitriadis, K. Georgiou, and A. Vakali, ''Social botomics: A systematic ensemble ML approach for explainable and multi-class bot detection,'' Appl. Sci., vol. 11, no. 21, p. 9857, Oct. 2021.
[155] A. Guerra-Manzanares, S. Nomm, and H. Bahsi, ''Towards the integration of a post-hoc interpretation step into the machine learning workflow for IoT botnet detection,'' in Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2019, pp. 1162–1169.
[156] P. P. Kundu, T. Truong-Huu, L. Chen, L. Zhou, and S. G. Teo, ''Detection and classification of botnet traffic using deep learning with model explanation,'' IEEE Trans. Dependable Secure Comput., early access, Jun. 15, 2022, doi: 10.1109/TDSC.2022.3183361.
[157] M. M. Alani, ''BotStop : Packet-based efficient and explainable IoT botnet detection using machine learning,'' Comput. Commun., vol. 193, pp. 53–62, Sep. 2022
[158] H. Bahsi, S. Nomm, and F. B. La Torre, ''Dimensionality reduction for machine learning based IoT botnet detection,'' in Proc. 15th Int. Conf. Control, Autom., Robot. Vis. (ICARCV), Nov. 2018, pp. 1857–1862, doi: 10.1109/ICARCV.2018.8581205
[159] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi, ''RTbust: Exploiting temporal patterns for botnet detection on Twitter,'' in Proc. 10th ACM Conf. Web Sci., New York, NY, USA, Jun. 2019, pp. 183–192, doi: 10.1145/3292522.3326015.
[160] H. Suryotrisongko, Y. Musashi, A. Tsuneda, and K. Sugitani, ''Robust botnet DGA detection: Blending XAI and OSINT for cyber threat intelligence sharing,'' IEEE Access, vol. 10, pp. 34613–34624, 2022
[161] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," in IEEE Access, vol. 7, pp. 41525-41550, 2019, doi: 10.1109/ACCESS.2019.2895334
[162] M. Melis, D. Maiorca, B. Biggio, G. Giacinto, and F. Roli, ''Explaining black-box Android malware detection,'' in Proc. 26th Eur. Signal Process. Conf. (EUSIPCO), Sep. 2018, pp. 524–528.
[163] J. Feichtner and S. Gruber, "Understanding privacy awareness in Android APP descriptions using deep learning,'' in Proc. 10th ACM Conf. Data Appl. Secur. Privacy, 2020, pp. 203–214.
[164] W. Guo, D. Mu, J. Xu, P. Su, G. Wang, and X. Xing, ''LEMNA: Explaining deep learning based security applications,'' in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., Oct. 2018, pp. 364–379
[165] A. Yan, Z. Chen, H. Zhang, L. Peng, Q. Yan, M. U. Hassan, C. Zhao, and B. Yang, ''Effective detection of mobile malware behavior based on explainable deep neural network,'' Neurocomputing, vol. 453, pp. 482–492, Sep. 2021, doi: 10.1016/j.neucom.2020.09.082
[166] G. Iadarola, F. Martinelli, F. Mercaldo, and A. Santone, ''Towards an interpretable deep learning model for mobile malware detection and family identification,'' Comput. Secur., vol. 105, Jun. 2021, Art. no. 102198, doi: 10.1016/j.cose.2021.102198.
[167] M. Kinkead, S. Millar, N. McLaughlin, and P. O'Kane, ''Towards explainable CNNs for Android malware detection,'' Proc. Comput. Sci., vol. 184, pp. 959–965, Jan. 2021, doi: 10.1016/j.procs.2021.03.118.
[168] M. M. Alani and A. I. Awad, ''PAIRED: An explainable lightweight Android malware detection system,'' IEEE Access, vol. 10, pp. 73214–73228, 2022, doi: 10.1109/ACCESS.2022.3189645.
[169] P. Barnard, N. Marchetti and L. A. DaSilva, "Robust Network Intrusion Detection Through Explainable Artificial Intelligence (XAI)," in IEEE Networking Letters, vol. 4, no. 3, pp. 167-171, Sept. 2022, doi: 10.1109/LNET.2022.3186589.
[170] Muna Al-Hawawreh, Nour Moustafa, "Explainable deep learning for attack intelligence and combating cyber–physical attacks," Ad Hoc Networks, Vol. 153, no. 103329, 2024, ISSN 1570-8705, https://doi.org/10.1016/j.adhoc.2023.103329.
[171] O. Arreche, T. R. Guntur, J. W. Roberts and M. Abdallah, "E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection," in IEEE Access, vol. 12, pp. 23954-23988, 2024, doi: 10.1109/ACCESS.2024.3365140.
[172] Arreche, Osvaldo, Tanish Guntur, and Mustafa Abdallah. "XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network Intrusion Detection Systems" Applied Sciences, Vol. 14, no. 10: 4170, 2024. https://doi.org/10.3390/app14104170
[173] C. I. Nwakanma, L. A. C. Ahakonye, T. Jun, J. M. Lee and D. -S. Kim, "Explainable SCADA-Edge Network Intrusion Detection System: Tree-LIME Approach," 2023 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Glasgow, United Kingdom, 2023, pp. 1-7, doi: 10.1109/SmartGridComm57358.2023.10333968.
[174] Larriva-Novo, Xavier, Carmen Sánchez-Zas, Víctor A. Villagrá, Andrés Marín-Lopez, and Julio Berrocal. "Leveraging Explainable Artificial Intelligence in Real-Time Cyberattack Identification: Intrusion Detection System Approach." *Applied Sciences* 13, no. 15 (2023): 8587.
[175] Yagiz, Muhammet Anil, and Polat Goktas. "LENS-XAI: Redefining Lightweight and Explainable Network Security through Knowledge Distillation and Variational Autoencoders for Scalable Intrusion Detection in Cybersecurity." *arXiv preprint arXiv:2501.00790* (2025).
[176] M. Gupta, C. Akiri, K. Aryal, E. Parker and L. Praharaj, "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy," in IEEE Access, vol. 11, pp. 80218-80245, 2023, doi: 10.1109/ACCESS.2023.3300381.
[177] Brandao, P. R., Mamede, H. S. & Correia, M. "Advanced Persistent Threats Campaigns and Attribution", Journal of Computer Science, Vol. 19, no. 8,2023, pp. 1015-1028. https://doi.org/10.3844/jcssp.2023.1015.1028

[178] Neupane, Subash, Ivan A. Fernandez, Sudip Mittal, and Shahram Rahimi. "Impacts and risk of generative ai technology on cyber defense." arXiv preprint arXiv:2306.13033, 2023

[179] Alexander Geiger, Dongyu Liu, Sarah Alnegheimish, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Tadgan: Time series anomaly detection using generative adversarial networks. In 2020 IEEE International Conference on Big Data (Big Data), pages 33–43. IEEE, 2020.

[180] Md Abul Bashar and Richi Nayak. Tanogan: Time series anomaly detection with generative adversarial networks. In 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pages 1778–1785. IEEE, 2020.

[181] Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In Artificial Neural Networks and Machine Learning–ICANN 2019: Text and Time Series: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17– 19, 2019, Proceedings, Part IV, pages 703–716. Springer, 2019

[182] Hossein Shirazi, Shashika R Muramudalige, Indrakshi Ray, and Anura P Jayasumana, "Improved phishing detection algorithms using adversarial autoencoder synthesized data", In 2020 ieee 45th conference on local computer networks (lcn), pages 24–32. IEEE, 2020.

[183] Briland Hitaj, Paolo Gasti, Giuseppe Ateniese, and Fernando Perez-Cruz. 2019. PassGAN: A Deep Learning Approach for Password Guessing. In Applied Cryptography and Network Security: 17th International Conference, ACNS 2019, Bogota, Colombia, June 5–7, 2019, Proceedings. Springer-Verlag, Berlin, Heidelberg, 217–237. https://doi.org/10.1007/978-3-030-21568-2_11

[184] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, Quoc Le, "LaMDA: Language Models for Dialog Applications", arXiv:2201.08239v3 [cs.CL], 2022

[185] " AI-Powered Talent & Sales Intelligence Platform | Draup", https://draup.com/draup-home/ (Accessed on : 22-12-2024)

[186] H. Trehan and F. Di Troia, "Fake malware generation using hmm and gan," in Silicon Valley Cybersecurity Conference, S.-Y. Chang, L. Bathen, F. Di Troia, T. H. Austin, and A. J. Nelson, Eds. Cham: Springer International Publishing, 2022, pp. 3–21.

[187] S. G. Selvaganapathy and S. Sadasivam, "Healthcare security: Usage of generative models for malware adversarial attacks and defense," in Communication and Intelligent Systems, H. Sharma, M. K. Gupta, G. S. Tomar, and W. Lipo, Eds. Singapore: Springer Singapore, 2021, pp. 885–897

[188] S. Sai, U. Yashvardhan, V. Chamola and B. Sikdar, "Generative AI for Cyber Security: Analyzing the Potential of ChatGPT, DALL-E, and Other Models for Enhancing the Security Space," in *IEEE Access*, vol. 12, pp. 53497-53516, 2024, doi: 10.1109/ACCESS.2024.3385107

[189] Bhatt, Manish, Sahana Chennabasappa, Cyrus Nikolaidis, Shengye Wan, Ivan Evtimov, Dominik Gabi, Daniel Song et al. "Purple llama cyberseceval: A secure coding benchmark for language models." *arXiv preprint arXiv:2312.04724* (2023).

[190] "Introducing llama: A foundational, 65-billion-parameter language model," https://ai.facebook.com/blog/large-language-model-llamameta-ai/, (Accessed on 12/22/2024).

[191] "Overview - advanced hunting | microsoft learn," https: //learn.microsoft.com/en-us/microsoft-365/security/defender/advancedhunting-overview?view=o365-worldwide, (Accessed on 06/22/2023).

[192] "Vulnerability Scanning Tools | Veracode", https://www.veracode.com/security/vulnerability-scanning-tools, Accessed on 12/22/2024

[193] "Yara - the pattern matching swiss knife for malware researchers," https: //virustotal.github.io/yara/, (Accessed on 012/22/2024)

[194] 4] "Azure openai service – advanced language models | microsoft azure," https://azure.microsoft.com/en-in/products/cognitive-services/openaiservice, (Accessed on 012/22/2024).

[195] O. D. Okey, E. U. Udo, R. L. Rosa, D. Z. Rodríguez, and J. H. Kleinschmidt, "Investigating chatgpt and cybersecurity: A perspective on topic modeling and sentiment analysis," Computers & Security, vol. 135, p. 103476, 2023. [Online]. Available: https: //www.sciencedirect.com/science/article/pii/S0167404823003863

[196] ["Bigid launches bigai, a 'privacy-by-design' llm designed to discover data," https://venturebeat.com/security/bigid-launches-bigai-a-privacyby-design-llm-designed-to-discover-data/, (Accessed on 12/22/2024).

[197] "Slashnext launches industry's first generative ai solution for email security," https://www.prnewswire.com/news-releases/slashnextlaunches-industrys-first-generative-ai-solution-for-email-security301757649.html, (Accessed on 12/22/2024)

[198] "How google cloud plans to supercharge security with generative ai | google cloud blog," https://cloud.google.com/blog/products/identitysecurity/rsa-google-cloud-security-ai-workbench-generative-ai, (Accessed on 06/22/2023)

[199] "Microsoft security copilot | microsoft security," https://www.microsoft. com/en-in/security/business/ai-machine-learning/microsoft-securitycopilot, (Accessed on 12/22/2024)

[200] Mansfield-Devine, Steve. "Creating security operations centres that work", Network Security, no. 5, pp.15-18, 2016

[201] Agyepong, Enoch, Yulia Cherdantseva, Philipp Reinecke, and Pete Burnap, "Challenges and Performance Metrics for Security Operations Center Analysts: A Systematic Review." Journal of Cyber Security Technology, Vol. 4, no. 3, pp. 125–152, 2019. doi:10.1080/23742917.2019.1698178.

[202] Chew, E., Swanson, M., Stine, K., Bartol, N., Brown, A., Robinson, W., 2008. Performance Measurement Guide for Information Security. Technical Report NIST Special Publication (SP) 800-55, Rev. 1. National Institute of Standards and Technology, Gaithersburg, MD. (https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-55r1.pdf) Last accessed on 2-9-2024.

[203] ISO/IEC 27004:2016, 2016. Information technology — Security techniques — Information security management — Monitoring, measurement, analysis and evaluation. Technical Report. International Organization for Standardization. https://www.iso.org/standard/64120.html. Last accessed on 2-9-2024.

[204] N. Salmi, "The present state of information security metrics," M.S. thesis, University of Jyväskylä, Jyväskylä, Finland, 2018.

[205] Vielberth, M., Böhm, F., Fichtinger, I., Pernul, G., "Security operations center: a systematic study and open challenges", IEEE Access 8, pp. 227756–227779, 2020, https://doi.org/10.1109/ACCESS.2020.3045514.

[206] Nathans, D., 2014. Designing and Building a Security Operations Center. Elsevier Science & Technology Books. Editor Steve Elliot.

[207] P. Keltanen, "Measuring outsourced Cyber Security Operations Center," M.S. thesis, South-Eastern Finland University of Applied Sciences, Mikkeli, Finland, 2019.

[208] Agyepong, E., Cherdantseva, Y., Reinecke, P., Burnap, P., 2020. Towards a framework for measuring the performance of a security operations center analyst. In: 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), pp. 1–8.

[209] Kokulu, F.B., Soneji, A., Bao, T., Shoshitaishvili, Y., Zhao, Z., Doupé, A., Ahn, G.J., 2019. Matched and mismatched SOCs: a qualitative study on security operations center issues. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, pp. 1955–1970.

[210] Onwubiko, C., 2015. Cyber security operations centre: security monitoring for protecting business and supporting cyber defense strategy. In: 2015 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), pp. 1–10.

[211] Alahmadi, B.A., Axon, L., Martinovic, I., 2022. 99% false positives: a qualitative study of soc analysts' perspectives on security alarms. In: 31st USENIX Security Symposium (USENIX Security 22), pp. 2783–2800.

[212] Crowley, C., Pescatore, J., 2019. Common and Best Practices for Security Operations Centers: Results of the 2019 SOC Survey. Technical Report. SANS Institute. https://www.sans.org/media/analyst-program/common-practices-securityoperations-centers-results-2019-soc-survey-39060.pdf.

[213] Ahlm, E., 2021. How to Build and Operate a Modern Security Operations Center. Technical Report. Gartner Inc.

[214] Logsign, 2020. Guide for security operations metrics. https://www.logsign.com/uploads/ Guide_for_Security_Operations_Metrics_Whitepaper_2f999f27cc.pdf.

[215] Simos, M., Dellinger, J., 2019. CISO series: lessons learned from the Microsoft SOC—part 1: organization. https://www.microsoft.com/security/blog/2019/02/21/ lessons-learned-from-the-microsoft-soc-part-1-organization/.

[216] Zimmerman, C., Crowley, C., 2019. Practical SOC metrics. https://www.fireeye.com/ content/dam/fireeye-www/summit/cds-2019/presentations/cds19-executive-s03bpractical-soc-metrics.pdf. fireEye Cyber Defense Summit 2019.

[217] Joonas Forsberg, Tapio Frantti, "Technical performance metrics of a security operations center", Computers & Security,Volume 135, 2023, 103529, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2023.103529.

[218] Huseyin Ahmetoglu, Resul Das, "A comprehensive review on detection of cyber-attacks: Data sets, methods, challenges, and future research directions", Internet of Things, Volume 20, 2022, 100615, ISSN 2542-6605, https://doi.org/10.1016/j.iot.2022.100615.

[219] Ahmed, Yussuf, Muhammad Ajmal Azad, and Taufiq Asyhari. 2024. "Rapid Forecasting of Cyber Events Using Machine Learning-Enabled Features" *Information* 15, no. 1: 36. https://doi.org/10.3390/info15010036

[220] Zaboli, Aydin, Seong Lok Choi, Tai-Jin Song, and Junho Hong. "A Novel Generative AI-Based Framework for Anomaly Detection in Multicast Messages in Smart Grid Communications." *arXiv preprint arXiv:2406.05472* (2024).

[221] De Diego, Isaac Martín, Ana R. Redondo, Rubén R. Fernández, Jorge Navarro, and Javier M. Moguerza. "General performance score for classification problems." *Applied Intelligence* 52, no. 10 (2022): 12049-12063.

[222] A. Kuppa and N.-A. Le-Khac, ''Adversarial XAI methods in cybersecurity,'' IEEE Trans. Inf. Forensics Security, vol. 16, pp. 4924–4938, 2021.

[223] J. Vadillo, R. Santana, and J. A. Lozano, "When and how to fool explainable models (and humans) with adversarial examples,'' 2021, arXiv:2107.01943.

[224] Loi, Michele, and Markus Christen. "Ethical frameworks for cybersecurity." *The Ethics of Cybersecurity* (2020):73-95.

[225] Kenneally, Erin, Michael Bailey, and Douglas Maughan. "A framework for understanding and applying ethical principles in network and security research." In *International Conference on Financial Cryptography and Data Security*, pp. 240-246. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[226] Macnish, Kevin, and Jeroen Van der Ham. "Ethics in cybersecurity research and practice." *Technology in society* 63 (2020): 101382.

[227] Mersinas, Konstantinos, and Maria Bada. "Behavior Change Approaches for Cyber Security and the Need for Ethics." In *The*

[228] *International Conference on Cybersecurity, Situational Awareness and Social Media*, pp. 107-129. Singapore: Springer Nature Singapore, 2023.

[229] Skinner, Tiffany, Jacqui Taylor, John Dale, and John McAlaney. "The development of intervention e-learning materials and implementation techniques for cyber-security behaviour change." ACM SIG CHI, 2018.

[230] Branley-Bell, Dawn, Lynne Coventry, Elizabeth Sillence, Sabina Magalini, Pasquale Mari, Aimilia Magkanaraki, and Kalliopi Anastasopoulou. "Your hospital needs you: Eliciting positive cybersecurity behaviours from healthcare staff." *Annals of Disaster Risk Sciences: ADRS* 3, no. 1 (2020): 0-0.

[231] Blythe, John. "Cyber security in the workplace: Understanding and promoting behaviour change." *Proceedings of CHItaly 2013 Doctoral Consortium* 1065 (2013): 92-101.

[232] Bustard, John D. "Improving student engagement in the study of professional ethics: Concepts and an example in cyber security." *Science and Engineering Ethics* 24 (2018): 683-698.

[233] Jagger, Suzy. "Affective learning and the classroom debate." *Innovations in Education and Teaching International* 50, no. 1 (2013): 38-50.

[234] Taylor, Carol, and Carol Robinson. "'What matters in the end is to act well': Student engagement and ethics." In *Understanding and developing student engagement*, pp. 161-175. Routledge, 2014.

[235] Tiwary, Devendra Kumar, and Uttar Pradesh. "Security and ethical issues in it: an organization's perspective." *International Journal of Enterprise Computing and Business Systems* 1, no. 2 (2011): 2230-8849.

[236] A. Dunmore, J. Jang-Jaccard, F. Sabrina and J. Kwak, "A Comprehensive Survey of Generative Adversarial Networks (GANs) in Cybersecurity Intrusion Detection," in *IEEE Access*, vol. 11, pp. 76071-76094, 2023, doi: 10.1109/ACCESS.2023.3296707.

[237] S. Bahadoripour, H. Karimipour, A. N. Jahromi and A. Islam, "An explainable multi-modal model for advanced cyber-attack detection in industrial control systems. Internet of Things, vol. 25, 101092, 2024. doi: https://doi.org/10.1016/j.iot.2024.101092

[238] I. H. Sarker, "Generative AI and Large Language Modeling in Cybersecurity. In AI-Driven Cybersecurity and Threat Intelligence: Cyber Automation, Intelligent Decision-Making and Explainability" pp. 79-99. Cham: Springer Nature Switzerland, 2024. Doi: https://doi.org/10.1007/978-3-031-54497-2

[239] A. Y. Wong, E. G. Chekole, M. Ochoa, and J. Zhou. "On the Security of Containers: Threat Modeling, Attack Analysis, and Mitigation Strategies. Computer Security. vol. 128, May 2023. doi: https://doi.org/10.1016/j.cose.2023.103140

[240] F. M. Teichmann and S. R. Boticiu, "An overview of the benefits, challenges, and legal aspects of penetration testing and red teaming. International Cybersecurity Law Review", vol. 4, no. 4, pp. 387-397, 2023.

[241] C. Chindrus, and C. F. Caruntu, "Securing the Network: A Red and Blue Cybersecurity Competition Case Study", Information, vol. 14, no. 11, pp. 587, 2023.

[242] M. R. Endsley, "Supporting Human-AI Teams: Transparency, explainability, and situation awareness", Computers in Human Behavior, vol. 140, pp. 107574, 2023.

[243] S. Hiremath, E. Shetty, A. J. Prakash, S. P. Sahoo, K. K. Patro, K. N. Rajesh and P. Pławiak, P, "A new approach to data analysis using machine learning for cybersecurity. Big Data and Cognitive Computing", vol. 7, no. 4, pp. 176, 2023.

[244] T. Arjunan, "Detecting Anomalies and Intrusions in Unstructured Cybersecurity Data Using Natural Language Processing", International Journal for Research in Applied Science and Engineering Technology, vol. 12, no. 9, pp. 10-22214, 2024.

[245] A. R. Al-Ghuwairi, Y. Sharrab, D. Al-Fraihat, M. AlElaimat, A. Alsarhan, and A. Algarni, "Intrusion detection in cloud computing based on time series anomalies utilizing machine learning", Journal of Cloud Computing, vol. 12, no. 1, pp. 127, 2023.

[246] Z. Liu, Y. Wang, F. Feng, Y. Liu, Z. Li, and Y. Shan, "A DDoS detection method based on feature engineering and machine learning in software-defined networks", Sensors, vol. 23, no. 13, pp. 6176, 2023.

[247] S. Al-Mansoori, and M. B. Salem, "The role of artificial intelligence and machine learning in shaping the future of cybersecurity: trends, applications, and ethical considerations", International Journal of Social Analytics, vol. 8, no. 9, pp. 1-16, 2023.

[248] N. Vemuri, N. Thaneeru and V. M. Tatikonda, "Securing trust: ethical considerations in AI for cybersecurity", Journal of Knowledge Learning and Science Technology, vol. 2, no. 2, pp. 167-175, ISSN: 2959-6386 (online), 2023.

[249] N. G. Camacho, "The Role of AI in Cybersecurity: Addressing Threats in the Digital Age", Journal of Artificial Intelligence General science (JAIGS), vol. 3, no. 1, pp. 143-154, ISSN: 3006-4023, 2023.

[250] Y. Kim, G. Dán and Q. Zhu, "Human-in-the-Loop Cyber Intrusion Detection Using Active Learning," in *IEEE Transactions on*

Information Forensics and Security, vol. 19, pp. 8658-8672, 2024, doi: 10.1109/TIFS.2024.3434647.

Dr. SHILPA ANKALAKI is currently working as an Assistant Professor, Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal. She holds a Ph.D. Degree in Computer Science and Engineering from Visveswaraya Technological University, Belagavi, India. She has authored several research articles published in various international journals and conferences. Her research interests include Machine Learning, Deep Learning, Data Mining and Artificial Intelligence.

Dr. Aparna Rajesh Atmakuri is an Associate Professor, Department of CSE, SoET, Centurion University of Technology and Management Bhubaneswar, Odisha. Her research interests include cybersecurity, cloud computing, IoT, and AI/ML. She has published several papers and book chapters in international conferences and journals, authored four technical books, and holds three patents.

Dr. Pallavi M received PhD in 2023, in Computer Science from Presidency University Bangalore, M. Tech in Computer Science &Engg. NMIT Bangalore, B.E degree in Computer Science &Engg. from Atria institute of technology Bangalore. She is currently working as an Assistant Professor in the Dept. of Computer Science & Engineering, Presidency University Bangalore, Karnataka. Her areas of interest include Machine Learning and Deep Learning.

GEETABAI S. HUKKERI received the integrated B.E. and Ph.D. degrees in computer science and engineering from Visvesvaraya Technological University, Belagavi, India. Since 2023, she has been an Assistant Professor with the Computer Science and Engineering Department, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India. She is the author of two books, four journal articles, and more than five conference publications. Her research interests include computer vision, information retrieval, big data, machine learning, and deep learning and its applications.

TONY JAN is currently the head of the School of IT and the director of the Artificial Intelligence Research Centre at Torrens University Australia. Tony was previously associate head and associate professor at the School of IT and Engineering at the Melbourne Institute of Technology and the University of Technology Sydney, respectively. Tony specializes in machine learning for cybersecurity and smart technologies, with over seventy articles in prestigious journals supported by several large research grants totaling over 20 million dollars in the domains of AI automation and homeland security.

GANESH R. NAIK ranked top 2% of researchers worldwide in Biomedical Engineering, is a leading expert in data science and biomedical signal processing. He received Ph.D. in Electronics Engineering, specializing in biomedical engineering and signal processing, from RMIT University, Melbourne, Australia, in December 2009. Currently, he is a senior lecturer in IT and CS at Torrens University, Adelaide, Australia. Previously, he was an academic and research theme co-lead at Flinders University's sleep institute. He held a Postdoctoral Research Fellow position at MARCS Institute, Western Sydney University, between July 2017 and July 2020. Before that, he held a Chancellor's Post-Doctoral Research Fellowship position in the Centre for Health Technologies, University of Technology Sydney (UTS), between February 2013 and June 2017. As a mid-career researcher, he has edited 15 books and authored around 160 papers in peer-reviewed journals and conferences. Ganesh is an associate editor for IEEE ACCESS, Frontiers in Neurorobotics, and two Springer journals. He is a Baden–Württemberg Scholarship recipient from Berufsakademie, Stuttgart, Germany (2006–2007). In 2010, he was awarded an ISSI overseas fellowship from Skilled Institute Victoria, Australia. Recently, he was awarded a BridgeTech industry fellowship from the Medical Research Future Fund, Govt of Australia.