Journal of Information Security and Applications 57 (2021) 102722

Contents lists available at ScienceDirect

Journal of Information Security and Applications

journal homepage: www.elsevier.com/locate/jisa

Weaponized AI for cyber attacks

Muhammad Mudassar Yamin [a,*], Mohib Ullah [a], Habib Ullah [b], Basel Katt [a]

[a] Norwegian University of Science and Technology, Norway
[b] University of Ha'il, Saudi Arabia

ARTICLE INFO

Keywords:
Artificial intelligence
Cybersecurity
Adversarial learning
Scenarios
Cyberattack
Cyber defense

ABSTRACT

Artificial intelligence (AI)-based technologies are actively used for purposes of cyber defense. With the passage of time and with decreasing complexity in implementing AI-based solutions, the usage of AI-based technologies for offensive purposes has begun to appear in the world. These attacks vary from tampering with medical images using adversarial machine learning for false identification of cancer to the generation of adversarial traffic signals for influencing the safety of autonomous vehicles. In this research, we investigated recent cyberattacks that utilize AI-based techniques and identified various mitigation strategies that are helpful in handling such attacks. Further, we identified existing methods and techniques that are used in executing AI-based cyberattacks and what probable future scenarios will be plausible to control such attacks by identifying existing trends in AI-based cyberattacks.

1. Introduction

In 2019, Burton et al. [1] described the Terminator as the benchmark set by Skynet in the Terminator movies and that people may be a few decades away from such self-aware AI. He terms such AI as "General AI". The researchers highlighted the warnings put forward by Henry Kissinger and the late Dr. Stephen Hawking of an impending AI arms race. The AI arms race is in full swing among countries such as China, Russia, and the US. The implications of the usage of arms use as well as its usage as part of cyber-security and protecting them from different threat actors must be of primary concern. Since AI is developed mainly by private companies, there is a lack of adequate regulation from countries; however, certain countries such as Canada, China, India, United Arab Emirates, United Kingdom, and the USA have taken new strides in this regard.

In strategic studies, the work on AI is not taken into consideration, particularly in military strategies and planning [1]. There are two thoughts in this domain: the utilization of AI in strategic studies would revolutionize military operations as well as revolutionize and benefit international security with better and more efficient decision-making solutions. Researchers [1] predict that a middle ground will help the operations side and it will be an evolutionary step in the human–machine AI decision-making aspect. They, explored the following four aspects: first, the AI that is being developed; second, its usage in cyberspace and cyber-security; third, its effect on combined military operations on air, land, and sea; and fourth, the strategic implications of AI for the deployment of weapons and the decision-making involved in their utilization.

The lack of AI in strategic planning stems from the fact that there is a lack of clear and concise knowledge in this specific domain. Numerous general technical terms and technological advancements are usually placed under the AI umbrella, which could be misleading. AI can be described as a technology with the human characteristics of thinking and analysis before taking actions. There are two types of AI: *Narrow AI* and *General AI*. Narrow AI can only perform a single task at a time. It has substantial usage as well as technological drawbacks. General AI is supposed to be able to achieve several tasks at a time; it is technologically advanced and future-based, which will be useful in strategic implications for military purposes. AI can also be a software/hardware mix, with technology at one end and subsequent hardware to support such technology on the other end. The definition of AI is based on the tasks and roles that it can perform, such as decision-making, military ops, lethal autonomous systems, etc. AI is a technology used in both civil and military ops. The New Zealand Navy is using it for logistics purposes. The NSA is using AI in its *PRISM* program to utilize *big data* for counterterrorism. Israel is using Al with the *Harpy drone* in dismantling enemy ops remotely, and China has developed a drone swarm technology that can be used to bypass enemy defenses [2].

1.1. The weaponization of AI

AI weaponization enables a more efficient use of conventional modes of weapons used in air, land, water, and space using AI-based decision-making. The weaponization of AI – particularly in nuclear

* Corresponding author.
E-mail address: muhammad.m.yamin@ntnu.no (M.M. Yamin).

https://doi.org/10.1016/j.jisa.2020.102722

Available online 5 January 2021

materials, toxins, and chemical materials – is documented and has also been considered in the context of climate manipulation and space usage [1]. AI weaponization in cyberspace is dangerous, as evident from the Microsoft bot "Tay" [3,4], which has exhibited sexist and racist tendencies when fed malicious data. If AI is fed with malicious data and weaponized with nuclear or other warheads, it could prove catastrophic. Therefore, there must be regulations in its strategic uses and implications. Further, as evident from the 2016 US Election [5,6], social media manipulation also showcases the adverse implications of unchecked and unregulated weaponized AI. We define weaponized AI as *malicious AI algorithms that can degrade the performance and disrupt the normal functions of benign AI algorithms, while providing technological edge attack scenarios in both cyberspace and physical spaces.*

Cyberspace is expanding at alarming rates. With such an exponential expansion, it is humanly impossible to check each bit of data individually. Thus, the utilization of AI is a core requirement for an efficient system. AI offers potential solutions and makes the security of cyberspace manageable within a reasonable period of time. Manipulation in AI software and hardware can result in drastic consequences. For example, researchers in UC Berkeley developed an autonomous driver fooling stop sign. If such a system would be weaponized with AI and military vehicles rely on it, people with nefarious intentions could hack into the system and change the armed vehicle's settings into classifying the signs in a wrong manner. This could lead to certain severe and unwarranted results. The IBM Research lab-created Deep locker malware and utilized it in the form of a WannaCry attack. The attack illustrates the vulnerability present in the systems and reveals how malicious AI can attack Internet-connected networks.

1.2. Weaponized AI in cyberspace

The malicious use of cyberspace with weaponized AI can be shown in two ways. The first is the integration in the current battle doctrine, and the second is integration in military operations in conjunction with quantum computing, big data, robotics, etc. The second method combines ops that are more realistic in the current scenario, as they can help in real-time analysis, enhance on-the-spot decision-making, maintain crucial chain of supply and demand, and do the dirty jobs in the military, which humans seldom do. All these would affect the power dynamics in conventional military ops. AI will bring the conflict to such a stage where "Mass" is essential. "Mass" is defined in the form of *Big Data* Analytics and not in the form of military might. The bigger the "Mass" of quantifiable data, the better the performance of AI. The interdependence of AI in various bodies of the military will increase with the further weaponization of AI.

The critical decision about weaponization of AI will depend upon humans. However, with the advent of time, an increasing amount of control and power will shift to AI; therefore, it must be kept in check and regulated as per requirements. In this sense, AI strategy, particularly in weaponized AI, is integral and significant and is a need of our time. All the aspects of AI, weaponized AI, and its implications in contemporary times must be discussed and investigated in order to develop adequate usage controls.

1.3. AI-powered vs classic cyberattacks

Cyber operations [7,8] can be conducted against specific kinds of individuals or organizations in a specific area using different kinds of cyber attacks. There are various kinds of cyberattacks but these attacks are not limited only to denial-of-service (DoS) and distributed denial-of-service (DDoS) attacks; man-in-the-middle (MitM) attack; phishing; spear-phishing attacks; injection attacks such as SQLI and XSS; jamming, eavesdropping attack; and malware attacks. These are considered as examples of the classical type of cyberattacks. In the present paper, researchers are focusing only on new types of cyberattacks powered by AI. The attacks that are being focused upon are mainly of three types:

- Data misclassification [9]
- Synthetic data generation [10,11]
- Data analysis [12,13]

AI is powered by data and our focus is to put on attack scenarios on audio, visual, and textual data. However, the technology behind these attacks can be applied to various other attack scenarios. *AI-powered vs classic cyberattacks* with respect to the well-known **STRIDE** [14] threat model is presented in Fig. 1.

In the present research, the researchers focus on identifying core technical concepts used in weaponizing AI for cyberattacks and the type of attacks that are currently being carried out by such weaponized AI. In the following sections, first, the background of the current research and related work is presented. Thereafter, technical insights from AI algorithms that are used in AI-powered cyber attacks are presented. Then, the current and future attack scenarios with relevant mitigation as well as defense strategies are provided; then, the article concludes with a discussion and conclusion.

2. Background

AI weapons are not a new concept; they can be traced back to the mid-1950s in the form of acoustic homing torpedoes. An immense amount of research and funding has made it possible to invent machines that take bold, decisive decisions to gain a military stronghold. DARPA [15] has been at the forefront of AI research and development. Nuclear-powered AI supersonic jets were a military thought once as well, so nothing is impossible for now. In the late 1980s, DARPA worked on autonomous weapons and they are now being used in military operations. Specifically, the aim was to see if plans of war using nuclear warheads can be controlled in real-time with the help of AI. "Survival Adaptive Planning Experiment" sought to see if AI, combined with human intelligence, could be used for militaristic operations. It was being tested with some success in 1991. The researchers in [16] argued that the most dangerous aspect of inculcating AI weapons might not be the actual AI-controlled weapons, but those controlled by both humans and AI, as those have a higher chance that things might go sideways. With the latest state of the art computing power, it is difficult to define what completely autonomous means. The difficulty in defining what is "an autonomous weapon" is at the forefront of future conflicts. For example, the US uses weapons like Phalanx fleet defense guns [17], which would be difficult not to ban in negotiations to control autonomous weapons. The sole disagreement on what can and cannot be defined as an autonomous weapon has led to initiate stalled efforts in the UN convention on conventional weapons for an outright ban on autonomous weapons.

In [18], researchers have argued about AI and how governments would attempt to use AI for their benefit or consider this a national security threat. The governments can either confiscate AI data or fund its research to outpace private AI researching firms. The researchers of this paper have indicated different cases where private firms are the first to utilize AI-driven technologies. However, continuous innovation in AI will lead to an arms race among different countries, which will result in more conflicts and violence, particularly with AI-powered weapons. International stability will be at risk if AI-powered technologies fell into the hands of a rogue government. An AI arm could be avoided effective implementation of the UN convention, global cooperation, more trade, better enforcement of non-proliferation treaties, more tolerance, and a lack of nationalistic sentiments. Achieving these goals would require substantial funds and support from different world bodies.

In 2016, Edward [16] argued that it is already too late to stop the shifting of military technologies to AI. He stated that instead of countering this technological advancement, it must be appropriately managed. He put the case forward that even if one country of military might install and implement the usage of AI in its military deployments, the remaining countries are likely to follow soon. He also believes that

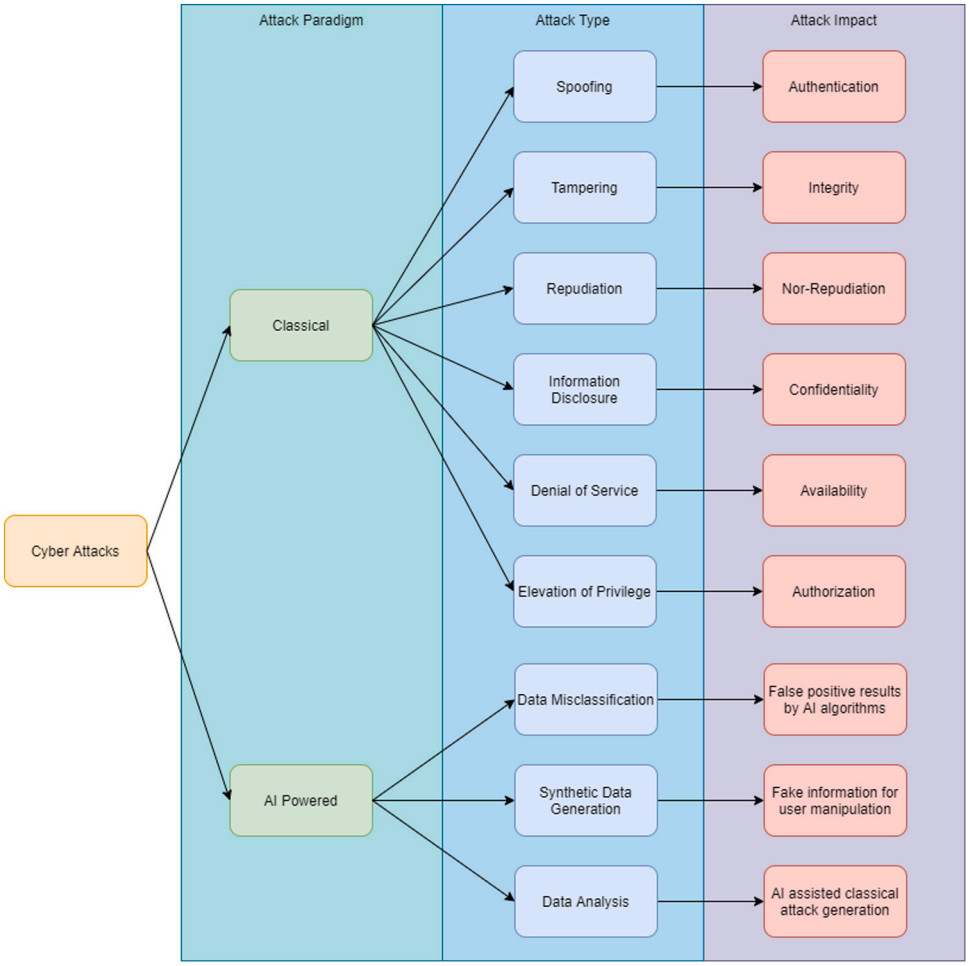

Fig. 1. AI-powered vs classical cyberattacks.

this stage has elapsed and we are truly at the brink of an AI arms race. He argued that instead of loosely checking a disastrous AI arms race, it must be managed appropriately and kept in check irrespective of any affiliations or interests that might deter the objective.

3. Related work

In 2018, Li [19] conducted a survey in which researchers identified the integration of AI with cybersecurity. Most of the work he reviewed was related to the detection of cyberattacks and the defensive usage of AI in the cyber domain. The researchers discussed the attacks employed against AI algorithms and shared defense mechanisms against them. The attacks on AI employ adversarial machine learning algorithms to induce noise in the input data of machine/deep learning algorithms to affect their classification process. The defensive measure consisted of modifying input data and modifying the neural network with additional classifiers to reduce its impact.

Further, in 2018, Duddu [20] surveyed the usage of adversarial machine learning in cyber warfare. He identified the threat model against machine learning algorithms and presented defense strategies against them. He identified seven adversarial machine learning attacks that are evasion attacks, poisoning attacks, equation solving attack, path finding attack, model inversion attack, blackbox attacks using transferability property and member inference attack. Moreover, he also identified vulnerabilities in machine learning algorithms that involve supervised, unsupervised, and reinforcement learning algorithms. The defense mechanisms he discussed were privacy-preserving AI algorithms and data pre-processing techniques to avoid attacks on machine learning-based algorithms.

Naveed et al. [21] conducted a survey in which they investigated the usage of AI in cybersecurity. The researchers identified the top contributors in the field with regard to their institution and countries. They used this information to develop a heat map to identify countries leading the research in the usage of AI for cybersecurity. The researchers identified that Chinese universities are leading the world in AI research related to cybersecurity and other prominent universities focusing on such research are located in Iran, Singapore, US, and India. Examples of the topics in which universities were conducting research were data mining, network traffic classification [22,23], anomaly detection [24,25], fraud detection [26,27], and adversarial machine learning [28,29]. The countries that are currently leading research on AI usage in cybersecurity are presented in Fig. 2.

The study of deep learning in computer vision reflects the adoption of deep learning methods in medical imaging. Various research articles have been published on this subject. The researchers in [30] have explored a recent breakthrough in deep learning—that is, generative adversarial network (GAN) and its medical imaging application. GAN is a specific type of neural network model where two networks are trained simultaneously, one focusing on image generation and another on discrimination. GAN is advantageous for domain shifts and in generating new image samples. The concept of adversarial training is rather standard now. The researchers discuss GAN and have explained its application in medical imaging. In addition, the researchers have employed various algorithms for this task and have also included the works of other researchers in this field. They have studied various GANs, their principles, and structural variants, and obtained a comprehensive view of medical imaging tasks using GAN. They also

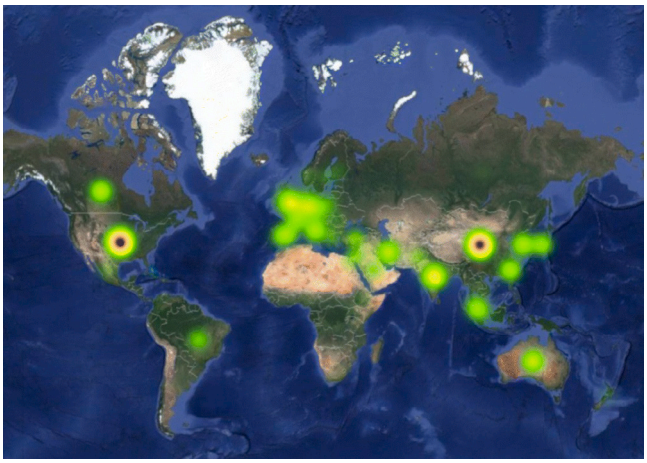

Fig. 2. Hotspots of AI research for cyber security [21].

categorized works according to canonical tasks, such as reconstruction, image synthesis, segmentation, classification, detection, and registration. GANs are being used in medical imaging in two different ways. The first method focuses on the generative aspect, and the other method focuses on the discrimination aspect.

Yi et al. [30] highlighted the usage of a generative adversarial network in cybersecurity. The researchers attempted to state the importance of these networks regarding medical imaging. GANs are notable due to data generation without explicitly modeling the probability density function. The adversarial loss due to discrimination paves the way for incorporating unlabeled samples into training and imposing higher-order consistency. It is beneficial for numerous cases, such as domain adaptation, data augmentation, and image-to-image translation. The following characteristics have urged researchers to link it with the medical imaging community and a quick adoption have been observed in numerous typical and new applications such as classification [31,32], image reconstruction [33], detection [34,35], segmentation [36], and cross-modality synthesis [37].

It is evident that the terms used in this field are new and the related work was not able to lay the foundation for this new domain. To fill this gap, the current work goes beyond the traditional perspective and attempts to address this type of research in a more holistic manner. There is a serious lack of multidisciplinary research regarding AI and its usage, particularly in the field of cyber security. Laws and ethical usage guidelines for AI-powered cyber adversaries are yet to be defined [12]; this has resulted in uncontrolled dual use and proliferation of AI technologies. Different countries are attempting to utilize such technologies as much as possible in order to take advantage over other countries. This initiated an AI arms race with no finish flag in sight. A multilateral approach at an international level to regulate the usage of such AI-powered technologies is the need of the hour.

4. Scope and methodology

As AI-based cyber attacks are relatively new, the researchers decided at an early stage not to follow the traditional survey methodology, including identifying literature based on keywords. For the survey, they have used the latest articles in the field – that is, from 2016 and onward – that contain the keywords "AI" and "cyberattacks" in their indexing terms. They considered the publication channels and the number of citations as a quality matrix to include the research articles. In addition, they focused on the applicability of the presented AI models in the cyber domain and their usage in cyber-security. First, they have identified the core technical concepts of the machine and deep learning network architectures that underlie the weaponizing of AI. Then, they divided the articles into three categories based on the data format – that is, audio, visual, and textual – and presented the scenarios according to this division.

5. Network architecture

When it comes to AI-based cyber attacks, GANs [38,39] play a pivotal role. Their main function lies in data generation (visual, textual, and audio) with applications including, but not limited to, video inpainting, audio synthesis, super-resolution, drug discovery, text mining, and synthesis of training data for training other deep networks, particularly in medical imagining where it is expensive to have a large amount of data. In a nutshell, GANs are essentially generative models that learn deterministic transformation T of data distribution p_s to approximate an unknown distribution p_n. From the architecture perspective, a GAN consists of two agents—that is, a generator and a discriminator. Essentially, both the generator and discriminator are deep networks but with a different task (different loss function). The generator network learns the distribution of the data and generates samples for the discriminator network. The discriminator is fed with two types of samples—one from the original data and the other generated by the generator network. The discriminator network's job is to classify if the generator sample originates from the original data or the generator network. During training, the generator attempts to produce more realistic samples to trick the discriminator, while the discriminator attempts to differentiate between the original and synthetic samples. GAN training is done end-to-end, and the network is assumed to be trained when it reaches the Nash equilibrium [40]. In the context of GAN, Nash equilibrium implies that when the loss of generator (failed attempts to fool the discriminator network) is equal to the loss of the discriminator network (lack of discrimination between the real and the synthetic sample). However, it is practically very difficult to establish such an equilibrium because the loss functions oscillate around the equilibrium position. Usually, after a few hundred epochs, the generated data is inspected visually or through an appropriate metric, and convergence is assumed if further training does not yield an improvement with respect to the selected performance metric. In Fig. 3, a functional schematic diagram of GAN is presented. Specifically, for the visual 2D data, the generator network G consists of encoder–decoder architecture. At the input, random noise is given to the network, and through convolution, pooling, and activation layers, the noise is transformed into an adversarial sample that could potentially be used in a variety of cyberattack scenarios. Contrary to the generator G, the discriminator network D takes the real and adversarial input and decides whether it is a real sample or an adversarial sample.

5.1. Loss functions

Loss function is the integral part of any deep learning algorithm. In case of the classical convolution neural network (CNN), the loss function is based on the cross entry for the classification. Mathematically, the standard cross entry can be written in the following manner:

$$\mathcal{L}_{Crossentropy} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\exp(W_{y_i}^T x_i + b_{y_i})}{\sum_{j=1}^{m}\exp(W_j^T x_i + b_j)} \tag{1}$$

where W and b are the weight matrix and the bias vector, respectively. M is the batch size, x_i is the data sample, and y_i is the true label of the data sample. In addition to the standard cross-entropy loss, a weighted average loss function can be introduced for different underlying problems. For example, a marginal loss factor [41] can be used in the loss function to maximize the margin between different classes in the data. Mathematically, such weighted average functions can be written in the following manner:

$$\mathcal{L}_{\mathcal{T}} = \mathcal{L}_{ML} + \gamma\mathcal{L}_{CE} \tag{2}$$

where $\mathcal{L}_{CE}$ is the cross-entropy and $\mathcal{L}_{ML}$ is the marginal loss. γ is considered the regularization parameter. The marginal loss $\mathcal{L}_{ML}$ is the

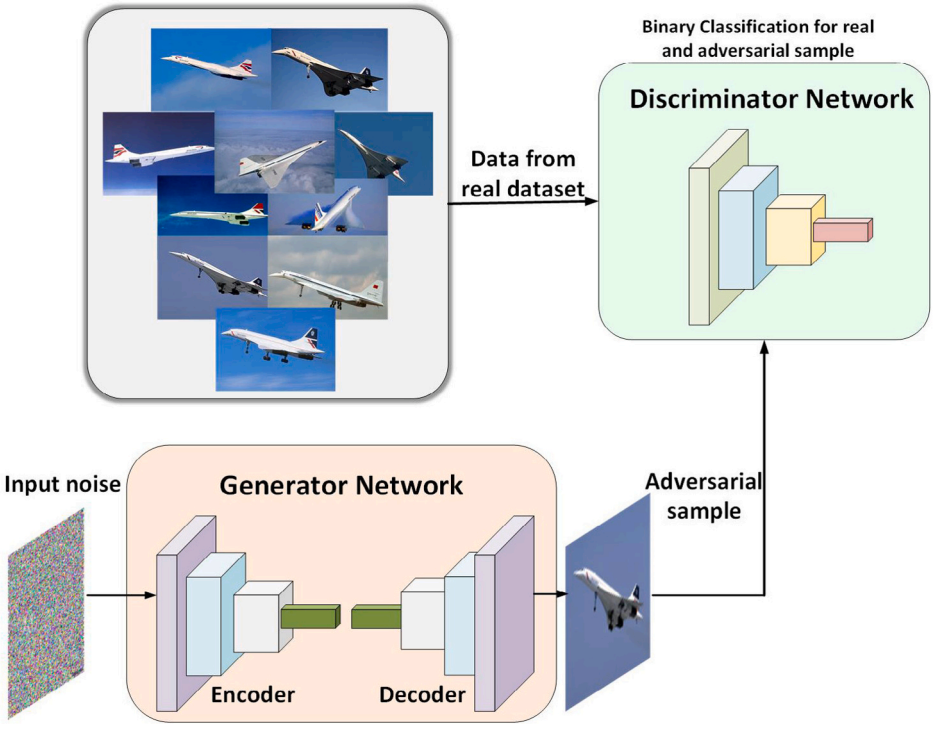

Fig. 3. GAN block diagram.

most important factor, as it maximizes the interclass distances and simultaneously minimize the intraclasses variations. In other words, the marginal loss forces the network to increase the margins among different classes. However, it attempts to reduce the distance among same classes. Numerically, the marginal loss $\mathcal{L}_{ML}$ is defined in the following manner:

$$\mathcal{L}_{ML} = \frac{1}{s^2 - s} \sum_{i,j,i \neq j}^{s} \left(\xi - \beta_{i,j} \left(\varphi - \left\| \frac{f_i}{\|f_i\|} - \frac{f_j}{\|f_j\|} \right\|_2^2 \right) \right) \tag{3}$$

where s is the batch size, φ is the threshold that separates the hyperplane of the classes, while ξ is the error margin among the hyperplanes. In addition, $\beta_{i,j}$ is a variable with two possible values—that is, either $+1$ or -1. For example, when f_i and f_j belong to the same class , $\beta_{i,j}$ is set to $+1$, and -1 otherwise.

In case of GAN, the generator G and discriminator D have their own loss functions. The task of optimization is to strict a balance between both losses and, mathematically, solve the following min–max problem.

$$\min_G \max_D \mathcal{L}(D, G) = \mathbb{E}_{x \sim p_{data}}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{4}$$

where p_{data} is the original distribution of the data and the p_z is the approximated distribution of the G network. Irrespective of the types of GAN, this is the classical min–max equation. Intrinsically, with such a formulation, the aim is to jointly optimize two parameterized networks – that is, the generator G and the discriminator D – and find an equilibrium between the two. During the optimization, the goal is to maximize the confusion of D while minimizing the failures of G. When such an equilibrium is achieved, the data distribution generated by the G yields similar distribution attributes as those of the D network. In order to see the distribution similarity, the most intuitive means is to calculate the distance between both distributions. For example, Kullback–Liebler divergence (KLD) can be used for this purpose. Mathematically, it can be expressed in the following manner:

$$KL(p_{data} \parallel p_z) = \sum_i^N p_{data_i} \log \frac{p_{data_i}}{p_{zi}} \tag{5}$$

While training GAN, we essentially increase the KL divergence and ideally, a score of 1 must be achieved, which corresponds to the exact same distributions. Having a good optimizer to tune the hyperparameter of the network is also of utmost importance. The most famous optimizers are briefly explained below.

5.1.1. RMS prop

Stochastic gradient descent is a well adopted and well studied gradient-based optimizer for deep learning models. However, classical stochastic gradient descent has three problems, which are

- It is slow
- It often gets stuck in the local minima
- It has an oscillatory behavior while converging to the solution

In order to address these issues, RMSprop [42] yields a better alternative for tuning the hyperparameters of the networks. The parameters are updated by introducing intermediate parameters I_{db}, and I_{dw}, which are also known as the running average of the squared gradients.

Let us assume that at iteration t, we compute the derivative of the parameter w as dw (convolutions filters) and of the parameter b (bias) as db, using stochastic gradient descent on the current batch. Then, the intermediate terms are calculated in the following manner:

$$I_{dw_t} = \gamma I_{dw_{t-1}} + (1 - \gamma) dw_t^2 \tag{6}$$

and

$$I_{db_t} = \gamma I_{db_{t-1}} + (1 - \gamma) db_t^2 \tag{7}$$

Once the intermediate terms are calculated, the main parameters are updated in the following manner:

$$w_{t+1} = w_t - \eta \frac{dw_t}{\sqrt{I_{dw_t}} + \varepsilon} \tag{8}$$

and

$$b_{t+1} = b_t - \eta \frac{db_t}{\sqrt{I_{db_t}} + \varepsilon} \tag{9}$$

For numeric stability, a small number ϵ is added in the denominator of both terms to avoid division by zero. The learning rate of η and γ is empirically selected and depends on the underlying application. With RMSprop, a larger learning rate can be used without the fear of divergence, as the intermediate terms I_{db} and I_{dw} keep the update in control.

5.1.2. Adam optimizer

Adaptive moment estimation (Adam optimizer) [43] combines the momentum and the RMSprop in the gradient descent, which essentially tunes the network learning toward faster convergence at each iteration. It has a number of hyperparameters (β_1, β_2, ϵ, η) that are empirically selected. In addition to the hyperparameters, in order to update the weights and bias, a few intermediate terms are calculated in each mini-batch that calculate the running average of the squared gradients (d_w, d_b, V_{dw}, V_{db}). Mathematically, it can be expressed in the following manner:

$$V_{dw} = \beta_1 V_{dw} + (1-\beta_1)dw \qquad V_{db} = \beta_1 V_{db} + (1-\beta_1)db \tag{10}$$

$$S_{dw} = \beta_2 S_{dw} + (1-\beta_2)dw^2 \qquad S_{db} = \beta_2 S_{db} + (1-\beta_2)db \tag{11}$$

At the t iteration, the intermediate parameters are updated in the following manner:

$$V_{dw}^{update} = V_{dw}/(1-\beta_1^t) \qquad V_{db}^{update} = V_{db}/(1-\beta_1^t) \tag{12}$$

$$S_{dw}^{update} = S_{dw}/(1-\beta_2^t) \qquad S_{db}^{update} = S_{db}/(1-\beta_2^t) \tag{13}$$

Finally, based on the updated intermediate parameters, the network weights and bias are updated in the following manner:

$$w_t = w_{t-1} - \eta \frac{V_{dw}^{update}}{\sqrt{S_{dw}^{update} + \epsilon}} \tag{14}$$

$$b_t = b_{t-1} - \eta \frac{V_{db}^{update}}{\sqrt{S_{db}^{update} + \epsilon}} \tag{15}$$

where η is the learning rate that is empirically selected.

5.2. Attention mechanism

Generally, deep networks are very susceptible to a small perturbation in the data and this attribute is heavily exploited in cyberattacks. In order to make the network resilient against such perturbation, an attention-based mechanism is introduced. Essentially, the attention mechanism is inspired by the human visual system. It is a relatively new term that is applied to deep models for improving the representation capability of the network and also helped the network to focus on the most important features and introducing robustness against a small perturbation in the data. In a nutshell, the attention mechanism improves the information flow among the layers of the network, which consequently helps in information accentuation or suppression and, as a result, yields a better representation for the underlying task. For a given set of feature maps $F \in \mathbb{R}^{C\times H\times W}$ by the network, the attention module extracts a 1D channel attention map $M_c \in \mathbb{R}^{C\times 1\times 1}$ and a 2D spatial attention map $M_c \in \mathbb{R}^{1\times H\times W}$. Mathematically, it can be represented in the following manner:

$$F_{channel} = M_c(F) \otimes F \tag{16}$$

$$F_{spatial} = M_s(F_{channel}) \otimes F_{channel} \tag{17}$$

where F is the set of feature maps given by a convolution neural network, $F_{channel}$ is the channel attention features maps, $F_{spatial}$ is the refined spatial attention feature map, and $\otimes$ indicates the elementwise multiplication.

5.2.1. Channel attention

The basic idea of channel attention is to ascertain the most important feature maps for a given input frame. Channel attention [44] utilizes average pooling and max-pooling for squeezing the spatial dimension of the input feature maps. The averaged pooled F_{avg}^c and max-pooled F_{max}^c descriptor are forwarded to a fully connected multilayer perceptron (MLP) with one hidden layer that generates the channel attention map $M_c \in \mathbb{R}^{C\times 1\times 1}$. The channel attention mechanism can be summarized in the following equations:

$$M_c(F) = \sigma(MLP(Avg_{pool}(F)) + MLP(Max_{pool}(F))) \tag{18}$$

$F \in \mathbb{R}^{C\times H\times W}$ is the feature map obtained through the network, while Avg_{pool} and Max_{pool} are the average and max pooling operations, respectively.

$$M_c(F) = \sigma(W_1(W_0(F_{avg}^c)) + W_1(W_0(F_{max}^c))) \tag{19}$$

Sigmoid function σ is used as the main activation function for the channel attention module, where $W_0 \in \mathbb{R}^{C/r\times C}$ and $W_1 \in \mathbb{R}^{C\times C/r}$ are the input to the hidden layer and the hidden layer is input to the output weight parameter for the multilayer perceptron (MLP). For retaining the parameters of the MLP, the hidden layer activation size is set to $\mathbb{R}^{C/r\times 1\times 1}$, with r as the reduction ratio.

5.2.2. Spatial attention

As compared to channel attention, spatial attention aims to localize the most informative aspect of the feature maps that is complementary to the channel attention. In order to calculate spatial attention, first, average pooling and max pooling operations are applied to the feature maps and then the resulting features maps are concatenated to obtain an efficient feature descriptor. In the resulting feature descriptor, a convolution layer is applied to generate the spatial attention map $M_s(F) \in \mathbb{R}^{H\times W}$. Mathematically, it can be defined in the following manner:

$$M_s(F) = \sigma(f^{9\times 9}([Avg_{pool}(F); Max_{pool}(F)])) \tag{20}$$

$$M_s(F) = \sigma(f^{9\times 9}([F_{avg}^s; F_{max}^s])) \tag{21}$$

where $F_{avg}^s \in \mathbb{R}^{1\times H\times W}$ and $F_{max}^s \in \mathbb{R}^{1\times H\times W}$ are the average and max pooling, respectively. The sigmoid function σ is used as the main activation function, while $f^{9\times 9}$ indicates the convolution operation with a filter size of 9×9. The refined feature map can be further processed to obtain the adversarial data sample for the cyberattack.

6. AI-based cyberattacks

In 2019, Chachra et al. [45] discussed machine learning strategies for identifying the relationship of such strategies to cybersecurity. Machine learning is based on algorithms that learn from previous experiences to understand reoccurring patterns instead of programming the patterns themselves. The need for machine learning lies in the fact that it proposes a second means to find a solution to such problems, like filtering out junk emails. Two types of algorithms can be applied in machine learning—supervised and non-supervised algorithms. A few methods are used to solve cybersecurity tasks by applying machine learning—regression, classification, clustering, association rule learning, and generative models. Applying regression in cybersecurity implies the detection of fraud or any other suspicious activity. Classification implies differentiating among a set of categories based on a training data set containing known category membership observations. Clustering is a method of unsupervised learning in which a technique of grouping of population or data points is employed. In comparison, generative models stimulate the actual data by creating a list of information parameters to test a specific application for injected vulnerabilities. Table 1 presents examples of researched malicious AI algorithms that are used to attack the classification of benign functionality of AI algorithms.

Table 1
Malicious AI algorithms used to tamper data for bypassing benign AI algorithm classifiers.

No.	Year	Target	Impact
1	2017	Traffic Signs [46]	Misclassification of the traffic sign by AI algorithms, which can lead to traffic accidents in autonomous cars
2	2018	Medical image data [47]	Misclassification of medical abnormalities by AI algorithms, which can lead to false diagnostics of health conditions
3	2018	Facial image data [48]	Misclassification of face images, which can lead to authentication bypass in certain scenarios
4	2019	Digital recommendation systems [49]	Data poisoning to AI algorithms, which results in wrong recommendations
5	2019	CT-Scan Data [9]	Misclassification of tampered CT-scan 3D images, which can lead to false diagnostics
6	2019	Speech audio data [50]	Adversarial attack on voice activate personal assistance, which can tamper their functionality
7	2020	Network intrusion detection systems [51]	Adversarial traffic generation to bypass the security of AI-powered network intrusion detection systems

Rege et al. [52] discussed how machine learning is being used for offensive and defensive purposes in the context of cyberattacks. The new world is becoming increasingly digital, with much interconnection among various technologies; this has been made possible due to network storage and sharing capabilities. However, with all this progress comes the risk of nefarious entities that want to sabotage the system for their gains. This is why cybersecurity is essential; in their research, the researchers examine offensive cybersecurity tools, such as botnets, spearfishing, and evasive malware. Further, the researchers also examined the defensive utilization of machine learning tools such as malware detection, which can be used for network risk scoring. The attacker needs to be correct one time, while the defensive mechanism must be accurate 100% of the time. A breach in cybersecurity can be devastating, as personal, confidential, and financial information are at the risk of compromise. With time, cyber intrusions are becoming more common and the complexity of these attacks is increasing day-by-day. Currently, cybersecurity has been enhanced using machine learning tools. Machine learning incorporates human behavior that duly creates data sets based on human behavior and then further recommends services based upon that behavior. Machine learning can prevent an attack before it can even take place using various threat detection protocols. For example, using data sets from the WannaCry [53] attack, machine learning can help mitigate attacks of the same nature in the future. Network risk scoring helps determine the most vulnerable parts of a network and what must be done to make these parts secure. In addition, machine learning can better predict them and lead toward a safer, more secure, and even more efficient network operation. Machine learning also helps to automate security-related tasks and increase human efficiency and response and analysis capabilities of the professionals concerned.

Shokri et al. [54] shared (1) how machine learning data sets can be compromised and their data be made viable to attack, (2) how data from such data sets can be leaked, and (3) what strategies can be utilized to mitigate its effects. Machine learning can be used to visualize data to make better sense of it. Companies such as Amazon and Google provide machine learning services [55] for clients to better understand and evaluate their data. The researchers in [54] put forward the idea of how interference attacks in these models can hamper their authenticity and credibility and what mitigating techniques must be utilized to combat them. The researchers created interference models and attempted to manipulate the data sets in different ways; they achieved 94% success in this regard. Success at such a level simply indicates that the data set is prone to vulnerabilities and attacks.

Table 2
AI algorithms used for synthetic data generation.

No.	Year	Name	Data type	Usage
1	2016	TextGAN [58]	Textual	Synthetic text generation through adversarial training
2	2017	FM-GAN [59]	Textual	Synthetic text generation through adversarial features
3	2017	MidiNet [60]	Audio	Synthetic audio generation
4	2017	Age-cGAN [61]	Visual	Face age predication with conditional generative adversarial networks
5	2017	CVAE-GAN [62]	Visual	Synthetic face image generation
6	2017	SenseGen [63]	Textual	Deep learning model for synthetic sensor data generation
7	2018	WGAN [64]	Visual	Synthetic brain MRI image generation
8	2018	ACGAN [65]	Visual	Synthetic liver medical image generation
9	2018	Pedestrian Synthesis GAN [66]	Visual	Synthetic pedestrian data generation
10	2018	HP-GAN [67]	Visual	Synthetic data generation for human motion prediction
11	2018	VAE-GAN [68]	Visual	Synthetic video generation from text
12	2018	WaveGAN [69]	Audio	Adversarial audio synthesis
13	2019	DermGAN [70]	Visual	Synthetic skin image generation
14	2019	CT-GAN [9]	Visual	Synthetic MRI medical image generation
15	2019	X2CT-GAN [71]	Visual	Synthetic X-RAY medical image generation
16	2020	D-NET [72]	Visual	Iris biometric data generation

The researchers that employed the Google prediction API [56], Amazon machine learning [57], hospital privacy data, and shadow training techniques concluded that leakage of data is a genuine possibility in machine learning data sets. The researchers devised a shadow training technique that is general but can help mitigate and counter such leakages and help dismember actual data from noisy interference data. They also showed how machine learning is playing a pivotal role in the fourth industrial revolution. Thus, this research is of prime and critical importance going forward into the future. Further, the researchers created a technique in which they feed the data into the model and in case of any interference or any malicious attempt, the model has the ability to detect such an attempt; in the case of medical data and privacy, this is of paramount importance for data preservation and privacy concerns. A few examples of identified attacks on machine and deep learning-based algorithms in the literature are illustrated below. Table 2 presents the unclassified AI algorithms that are being utilized for the generation of synthetic data sets. These algorithms can be used to develop new and more robust AI or they can be used to develop fake data and information.

On the classical offensive cyber security operations end of the spectrum, machine learning can be used to gain unauthorized access into networks to retrieve data and information that can be used to create evasive malware and damage the infrastructure it comes across.

Table 3
AI-powered tools that use data analysis for offensive cyber operations.

No.	Year	Name	Usage
1	2017	DeepHack [73]	AI-powered tool to generate injection attack patterns for database applications
2	2018	DeepLocker [74]	AI-powered tool that emulates an APT for launching complex cyber attacks
3	2018	GyoiThon [75]	AI-powered tool for information gathering and automatic exploitation
4	2018	EagleEye [76]	AI-powered tool for social media information reconnaissance using facial recognition algorithms
5	2018	Malware-GAN [77]	AI-powered tool used for generation of malware that can bypass security detection mechanisms
6	2019	uriDeep [78]	AI-powered tool that generate fake domains for usage in different attack scenarios
7	2019	Deep Exploit [79]	AI-powered tool that automates Metasploit for information gathering, scanning, exploitation and post exploitation
8	2019	DeepGenerator [80]	AI-powered tool to generate injection attack patterns for web applications

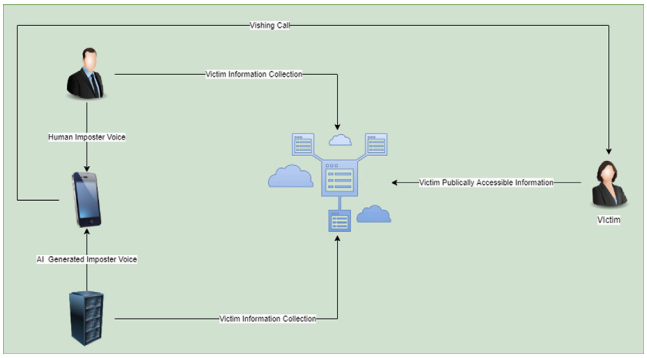

Fig. 4. Vishing attack scenario.

It can also be used as a spear-phishing tool to obtain data from selected individuals for various purposes. On the flip side, cyber-criminals can avoid detection and carry on with their nefarious plans by feeding false training data into the machine learning model, altering the machine learning algorithm's model code. The worst of all, evade detection through the latest technology, which can prove extraordinarily tenacious and dangerous. Table 3 presents AI-powered tools that are used for offensive cyber security operations.

6.1. Audio

In a recent research article [50], the researchers demonstrated the capability of jamming the functionality of audio-activated devices. These devices include Apple Siri and Google assistant, which use a wake-up word to activate them. Researchers used GANs to generate unusual noise, which can be used to disable the appropriate functionalities of devices. In another type of cyberattack, attackers use a technique called vishing [81], in which they use telephone communication as an attack vector and call the vulnerable victim while posing as someone else.

The typical attack scenario is presented in Fig. 4 in which the attacker is posing as a CEO of a company to obtain certain information or execute an action on the company employer with fake voice synthesis techniques [82]. The scenario is now entirely changed, and the attacker

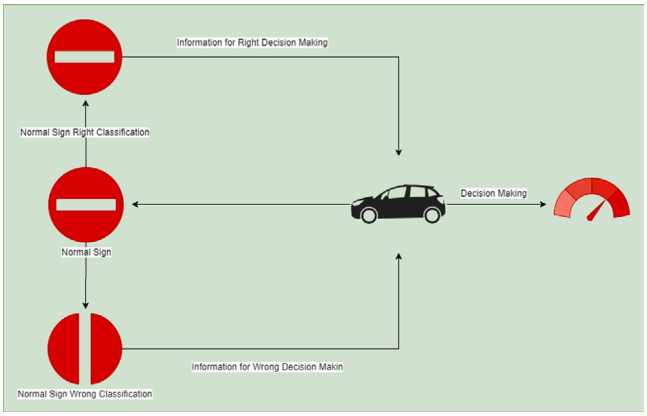

Fig. 5. Autonomous car driving attack scenario.

does not need a voice actor to perform the required action. The attacker can use a deep fake [83] voice generator [84] to launch a complex series of attacks targeting multiple individuals simultaneously using a real voice. Hence, vishing becomes more like a spear-phishing attack.

6.2. Visual

In 2017, Papernot et al. [46] put forward various techniques and methods in which adversarial attacks can bypass machine learning models and their deep neural networks. They examined how an adversarial network can infiltrate and attack the network from within. They examined the models hosted by Amazon and Google and attacked their Deep Neural Networks (DNNs). Further, the researchers also found that this kind of black-box strategy attacks can help in crafting a network that can be adaptive in the future and resist even more strategies that might be used in their defense.

A machine learning classifier can be defined by a set of input and expected output. The input and output would use a system of classes for accurately classifying a single input into a defined output class. The researchers used it on a stop sign where the machine learning classifier misclassified the stop sign by manipulating the stop sign image by adding noise generated from the adversarial machine learning algorithm. Such manipulation on a large scale can prove to be rather problematic. The simple depiction of the scenario is provided in Fig. 5.

Through this research, the researchers created an algorithm based on fabricated data to root out and counter misclassified DNN data. This research is of paramount importance when considering adversarial attacks on DNN. The researcher attacked a DNN hosted by MetaMind [85] and found that 84.24% miss-classification occurred due to the adversarial samples. The researchers also calibrated their machine learning models by comparing it with Amazon and Google's hosted classifiers and found rates of success of 96.19% and 88.94%, respectively. The researchers also found that their mode of attack is resilient to a string of defensive mechanisms. Further, the researchers also formalized and put forward an intuition-based adversarial sample as well.

In another work, Mirsky et al. [9] examined how using deep learning and an AI attacker can breach medical records for various nefarious purposes, whether for political or monetary purposes. The researchers presented a generative adversarial network for 3D images for CT scans that they called CT-GAN. Data on medical imagery can be manipulated for political, research sabotage, ransomware, or revenge purposes. It can be done by falsifying medical records – for example, by removing or adding cancers, tumors, etc. – to fulfill a hidden agenda. MRI and CT scans are used to produce 3D images of a body. MRI uses magnetic fields, whereas CT scans rely on X-rays to view different body images for various purposes. The scans are taken through a central system and stored and retrieved when required. They are susceptible to attacks

because, in health care machinery, data security is not given necessary importance as data privacy. The researchers examine how lung cancer data from CT scans can be manipulated and how to mitigate it. They used GANs, in which there are two neural networks, where one is a generator and the other is a discriminator. The generator creates fake images, and the discriminator differentiates the real from the fake images. For their attack, the researchers used their own developed CT-GAN for the tampering of medical imaging. In the attack model, the researchers investigated the following questions:

1. How can the supposed attacker can inflict and encompass fake imagery ?
2. How is such an attack implemented and how is the image doctored ?
3. What countermeasures can be utilized to combat this threat ?

Three radiologists with 2, 5, and 7 years of experience, along with an AI algorithm, were used to test the sample with tampered images in two ways:

- In blind trials, in which the radiologists were given 80 CT scans and were asked to review them.
- In open trails, in which they were given 20 CT scans and informed that some were fake and doctored and some were real and were asked to find out which one is real and which one is fake.

In the blind trials, the results were very alarming. One can rely on the diagnosis of the radiologists; they were very accurate with their diagnosis. However, the AI could not classify doctored images accurately in numerous cases, which is extremely alarming as numerous radiologists utilize AI in their diagnosis. The radiologist treated the injected tumors and cancers as normal—that is, they did not report anything abnormal. However, what is more distressing is the fact that one-third of the injected cancers required immediate medical treatment as per the radiologist. In addition, all the injected cancers required follow-ups and referrals as per the radiologists.

Further, according to the open trials, the radiologists could not find the difference between the real or fake cancers, where the cancerous cells have been added to eliminate eluded sight. With the knowledge that there may be abnormalities, the radiologists could find more instances of doctored imagery. The researchers suggested to increase data security, antivirus software on medical devices, and digital watermarking on medical images as a security countermeasure for such attacks. It was discovered that the framework fooled both the radiologists and AI.

6.3. Textual

Machine learning is susceptible to attacks and is vulnerable to external interference. Researchers in [13] used pdf files and checked whether malware attacks on such files are identified by the machine learning apparatus and how they can be better dealt with. A framework is proposed to exploit and find loopholes and vulnerabilities and this resulted in revealing security lapses in the system. For adversarial action purposes, machine learning has already been used for its capabilities to combat spam and malware. However, what happens in the case of an adaptive adversary that can pull off manipulative tasks and breach the safety and security of the data at hand ?. In this case, proactive adversarial actions are required, which examine the vulnerabilities present, effects of attacks, and mitigation strategies for the attacks. They identified different ways in which the attacks can be evaded or the data be made more secure and how the adversarial machine learning apparatus can be made proactive. In their scenario, they utilized the min–max approach, which was the Nash equilibrium approach to attain the goals. Researchers were able to identify a few keywords that are easily manipulated in pdf files that includes /open action, /comment, /root, and /page layout. These keywords infiltrate safe pdfs with malicious software and a tampered pdf that can be used in attacks. The researchers proposed ways that enable them to evade machine learning-based attack detection mechanisms. For example, they explored pdf files with different data sets to modify their attacks in order to evade detection.

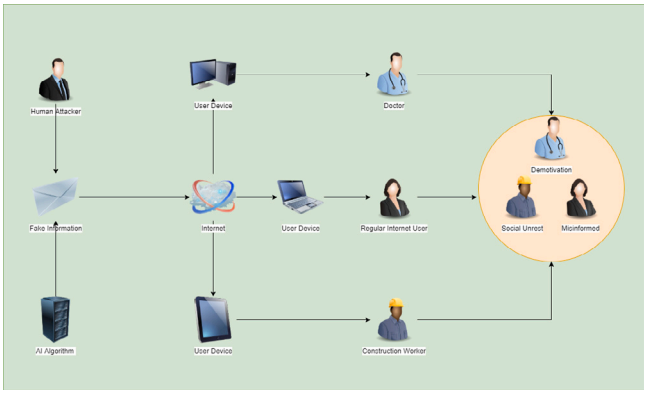

Fig. 6. Fake news generation and its impact on AI.

Numerous researchers are working on the generation of fake text to create content for websites. However, this content generation process can be used for the generation of fake news well. We are demonstrating one of the examples of fake text generation [10,11].

Input Text: *"Fake news generated by AI is"*

Output Result:

> Fake news generated by AI is icky, but it appears likely that at least some of the attacks will be fake. This can be made less of an issue by reducing information overload in an article by only using links that really say something.
> If you feel like you should put a bit of effort behind your blog or article without the intention to generate clickbait for the media, you should consider creating a more intelligent post — one that cannot be manipulated based on data, but can easily be corrected or modified. That way, people can trust that your site is truly original.
> Conclusion:
> A lot of people think that we are at a breaking point in the debate about AI and that it will have a massive impact on businesses and life in general. However, it is crucial that we maintain a balance between the "human factor" and machine learning.
> AI is definitely on the way to becoming an important part of the digital economy and our life, but not by a long shot. The more you can minimize the risk, the better it can be. We all do things for a reason.

In a recent case, an AI-powered bot went unnoticed for a week on reddit [86]. It was communicating with humans by posting comments with synthetically generated data and was able to convince a few people that it was a real human. This example makes it evident that such technology can be used for generating and spreading of content that may cause social unrest. An attack scenario representing such malicious use of AI is presented in Fig. 6.

7. Expected future scenarios

AI encompasses over 50 algorithms, and selecting the suitable AI algorithm for our needs is based upon problem requirements and in-depth knowledge of the field. An IBM machine defeated the world chess

champion Gary Kasparov in 1997. The IBM Watson [87] (an AI-based machine) defeated two human champions in 2011, while google AlphaGO defeated the world GO champion in 2016 [88]. The researcher in [89] states that if the AI went against the AI, it would most probably result in a draw. IBM Watson's success leads to the utilization of AI in the medical field and the diagnosis of various diseases. Currently, X-rays, genetic tests, and retina scans can be performed using AI. The researchers in [89] argued that unless and until we have more data, we have to keep doctors in the loop and not rely entirely on AI in the medical field. AI can prove to be monumental for the medical field, business, and communication and help to accelerate the fourth industrial revolution. Further, AI can be used on various ways for various purposes.

1. Through implants, AI-based robots can provide entertainment and pleasure [90].
2. Finding the most feasible, market competitive rates from among all the resources available (offers, discounts, etc.) [91].
3. AI can be used to morph images of one individual onto another and detect these [92].
4. AI can be used in developing autonomous combat vehicles [93].

AI can be used for both defensive cybersecurity as well as offensive attacks. Further, AI can be used against or in support of another AI in cyberspace. An offensive AI might penetrate a vulnerability and cause harm. A defensive AI might cover up the vulnerability and secure the network. One such example is evident from the DARPA cyber grand challenge [94], in which AI-based cyber attackers and defenders competed against each other. Adversarial machine learning and GAN, as utilized in DNNs, can be used in the offensive and defensive AI cyberspace. The greatest challenge for AI is obtaining the necessary accurate data to create accurate models. In the future, with the latest technology and better communication infrastructure with 5G/6G capabilities [95], AI can be used to solve natural and biological problems more efficiently. Machines can learn in real-time, make effective decisions, and lead to a more efficient learning system. The fast advent of AI can potentially lead to an AI arms race, which will be detrimental to international security and stability and lead to an increase in cyberattacks. The US passed the Algorithm Accountability Act [96] to make AI developers accountable. The EU passed legislation for a safer digital cyberspace [97]. In addition, IEEE's AI/AS initiative [98] is focused on developing AI-based technologies that have a productive use for humanity.

In 2018, Falco et al. [99] proposed that smart cities comprise the Industrial Internet of Things (IoT) and encompasses an industrial control system that helped in addition to the IoT infrastructure. The researchers proposed that the smart city infrastructure is prone to cyberattacks because local governments are neither involved nor educated in the domain of cybersecurity. Moreover, the policy and discussion related to cybersecurity is not highlighted at the government's policy and decision-making levels. The researchers discussed various attack capabilities by modeling attack trees in which the cybersecurity infrastructure can be compromised. They proposed an attack model and discussed mitigation against it to avoid catastrophic consequences. In an interconnected world where data is of paramount importance and is comparable to utilities – such as water supply and electricity – rely on it. Further, machine learning and AI can enhance their quality of service using the right data. However, vulnerabilities and infringements are present and not all are depending on the Internet; a few of these can infiltrate via the physical medium as well [100].

Ransomware attacks against government bodies are becoming increasingly common and must be countered effectively and efficiently. The deployment of such an attack methodology in an isolated environment will help us understand their effects and consequences. Consequently, this can help in proposing the appropriate mitigation techniques to secure systems and save taxpayer money. Smart cities can be made even safer and more secure using publicly available tools; however, they can also be used for malicious purposes by qualified entities. Excessive information exposure in cybersecurity has led to much interest from both benign and malicious users in this domain, as AI is enshrined in numerous critical government offices. In a complex industrial setting, the deployment and execution of potential attack methods in a safe sandbox enable us to understand and model the means through which these systems can be attacked and their security be penetrated. This knowledge and experience are important to be able to examine the appropriate loopholes and, thus, they are necessary for the safe usage of industrial systems.

In conclusion, this attack model is based on CCTV cameras and it is not conclusive whether it can be scaled toward the industry side. However, this model makes big data and cyberspace more secure. A CCTV-based attack model depends on the underlying application. For example, one of the possible scenarios is the manipulation of traffic flow [101,102]. Similarly, manipulating the crowd flow [103,104] to generate a potential stampede in a public gathering, diverting people flow [105,106] through tracking to spark overcrowding, and mixing the normal flow of people [107,108] with other traffic entities are some of the examples of such attacks that may take place in ever-evolving future smart cities [109]. One key factor that is commonly limiting AI research utilization is reproducibility. Strict protocols on data collection and reproducibility are required for the further advent of AI research.

8. Mitigation strategies

Computing devices and the internet have become an integral part of our society, and security professionals must adopt comprehensive strategies to counter the threats posed by AI-powered attacks. From this perspective, the researchers in [45] presented methods used to counter cyber security attacks and their relationship with machine learning models. The researchers have identified pitfalls and characterize challenges regarding machine learning algorithms in cybersecurity attacks. The researchers employed three basic network security systems for countering external and internal security breaches—namely, signature-based, anomaly-based, and hybrid. The signature-based detection technique refers to the detection of known attacks with those attacks' signature. Anomaly-based strategies observe the regular network and machine behavior and detect abnormalities as deviations from normal behavior. Lastly, hybrid detection is a combination of signature and anomaly detection [110]. It finds the detection rate of known intrusion and lessens the fake array of unknown attacks. Given below is the technical background for tackling AI-powered cyber attacks and a few examples of usage in the field.

8.1. Regularization as a mitigation tool

One of the classical problems in any deep learning algorithm, including GANs, is overfitting, which implies that the underlying algorithm learns the distribution of the training data so well that it performs almost perfectly on the training set but when it comes to the validation set, the results severely deteriorate. In the classical deep learning algorithms – such as the convolutional neural network, auto encoder, and boltzmann machine – overfitting is relatively easier to analyze due to the relative simpler architecture of such algorithms. However, when it comes to GANs, it becomes rather challenging to address the overfitting problem. In the realm of cyberattacks on the deep learning algorithms, particularly GANs, hackers mainly exploit this attribute of the learning and attempt to generate data to foul the networks. There are different strategies to address this limitation. In a nutshell, they are known as network regularization; in the last few years, developing sophisticated regularizers is one of the mainstream topics in the machine learning research community. Mathematically, a machine learning algorithm approximates a function $f_w : x \rightarrow y$ with trainable weights $w \in W$. The function f_w depends on the underlying network. For example, in

the case of a convolutional neural network, auto encoder, etc. f_w is a parametric function with the parameters w and b. In the case of GANs, f_w consists of two components—discriminator and generations. Hence, the function f_w can be defined in different ways depending on the architecture and the functionality of the application dependent network. Irrespective of the network architecture, training is essentially an optimization scheme where the aim is to identify a weight configuration w^* that yields the minimum error/loss for the given loss function $\mathcal{L} : W \rightarrow \mathbb{R}$, where $W \in \mathbb{R}$ represents the weight parameters belonging to real numbers. The minimization problem can be written in the following manner:

$$w^* = \min_w \mathcal{L} \tag{22}$$

Usually, the loss function takes the form of expected risk:

$$\mathcal{L} = \mathbb{E}_{x \sim p_{data}}[E(f_w(x), y) + R], \tag{23}$$

where $f_w(x)$ is the predicted value by the network and y is the true label if the classification problem is considered. Hence, in such a setup, the loss function can be divided into two discrete components—that is,

- Error function E
- Regularization R.

The error component E mainly depends on the true label of the data and the predicted label by the network depending on considered distance metric. However, the regularization term penalizes the model based on other criteria. It may depend on anything except true labels, for example, on the learnable hyperparameters of the network. In terms of implementation, regularization is incorporated in the network through the L1/L2 norm, dropout, early stopping, and a variety of data augmentation strategies. In terms of cyberattacks, we argued that a model with better generalization capabilities is likely to handle such attacks better. Hence, replying on a generalized model, the economic consequences can be substantially mitigated.

8.2. AI-model security threats

Securing the AI classification models is rather important for their expected usage. AI models are basically prone to three types of attacks:

1. Evasion
   In evasion, the input provided to AI algorithm is specifically tampered that enables them to bypass the right classification mechanism [9].
2. Poisoning
   In poisoning, the training data supplied to the AI classifier is altered which effect the classification of AI algorithm [4].
3. Stealing
   In stealing, the AI algorithm input and output is analyzed to identify the model properties and develop an own model to counter those properties [46].

Deep-pawn [111] was developed in 2016 and can be used to test the above types of attacks types on AI models. It is imperative that researchers test their developed models in controlled environments [112] for different input and output scenarios before deploying them in the real world.

9. The use of AI in cyber defense

9.1. Anti-phishing

Phishing is a method employed to steal personal data via the Internet and is a serious cyber-crime. With the passage of time, the number of phishing attacks has been considerably increasing. Phishing is detected by the blacklisting method in which a comparison is made with the phishing links that have been earlier reported by victims. Another method of phishing detection, like image analysis, is time-consuming. Various machine learning algorithms have been developed for phishing detection. Over time, the researchers have put efforts into the machine learning algorithm to train phishing detection models.

In their research, Chen et al. [113] dealt with the subject of a phishing attack. The researchers developed a new AI-based technique for the detection of phishing attacks. They extracted 28 features to train machine learning algorithms. The researchers employed a new approach by using the SMOTE algorithm for phishing detection. They have used the extracted features for the feature evaluation module. The three employed methods of phishing detection were web page similarity, blacklisting, and image processing. These methods were divided into three categories—namely, blacklisting, visual similarity analysis, and heuristic analysis. The researchers employed the data processing layer to generate data by the SMOTE method and analyze it via ANOVA, X2, and *Information Gain* method. The researchers extracted the training database by collecting phishing sites and by generating their URLs in the pre-processing unit. With the feature extraction module, the researchers performed feature extraction. After the URL was sent to the feature evaluation module, the unnecessary information was deleted. The three bases of assessment were data cluster, distribution, and independence. The researchers assessed the difference and accuracy of SMOTE and recommended an advanced framework for phishing research.

9.2. Use of AI in cyber kill chain

Cyber attacks are increasing rapidly with the advent of time. However, a human-centric response is not as prompt to counter such attacks. Chomiak et al. [114] conducted a study in which he argued that AI could be used along the cyber kill chain [115] to resist and counter possible cyberattacks. The researchers first examined AI's capabilities in the following domain: the kind of knowledge that is required to counter the attack, mimicking perception based on human nature, and finally, the capability to make on-the-spot decisions. Based on these aspects, it was then asserted at which point in the kill chain the capabilities of AI are required the most; further, it was examined where AI was most efficient—whether it was the reconnaissance of the system, the intrusion into it, or the exfiltration of data. The researchers asserted these through their research, as cyber-security is becoming an increasingly significant field of study with more sophisticated attacks and dangerous threat patterns with time. Thus, AI, coupled with human intelligence, can prove to be a game-changer in the realm of cyber-security.

The researchers identified that AI is here to stay, and with time, the AI models and the neural networks associated with them will get better and more efficient along the way. The efficiency of the AI networks depends upon the data that is being fed into it. The researchers argue that AI will play a vital role in the future of reconnaissance, infiltration, data exfiltration, and privilege escalation, and they would predict the advent of an increasing number of cybersecurity tools. This leads to further discoveries and inroads into the manner in which AI can benefit society as well. One drawback that the researchers identified is that AI is not entirely autonomous. Currently, it requires a well-trained professional to operate correctly. Data privacy is also a key issue when dealing with AI. When AI becomes completely autonomous, there is a lack of a legal and regulatory framework to keep it in check [12]. The researchers summarized that the AI hype has not met its reality yet, and that AI is not yet completely autonomous. There must be an effective bridge between man and machine for this AI system to work coherently, efficiently, and securely.

9.3. Cyber-attack visualization

Sundarara [116] suggested that it becomes incredibly monotonous to respond to an attack with human input in a smart grid system. In an automated attack, the response is not timely and is divisive to counter

the attack. The researcher has put forward ways in which the automated response to attacks can be improved and made more efficient and effective. They propose a framework based upon Kafka, Apache Spark, R, and neural network to secure the grid system in the case of an automated attack. Attacks on smart grid systems are on the rise, as evident from the attacks on Ukraine's smart grid system and the power plants of Iran. With an attack on cyberspace and a direct effect on the physical space and the advent of related technologies, the grid systems have become increasingly vulnerable; this is why an effective countermeasure is required in a nefarious attack. Currently, most responses to such attacks are made through the decisions of human operators, which is both stressful and erroneous and can lead to catastrophic results. Their research put forward a trimodular framework in which human input collaborates with cyber tools to combat such grid attacks. This does not change the existing cyber grid infrastructure, but complements it and places an added layer of security.

The researchers [116] also conducted a case study in which they took three different places in Florida and took their photovoltaic (PV) tied grid system. The load and power patterns were set, and the settings were determined according to the needs of the locality; an attacker was introduced into the system to look into the system and infiltrate it, thereby causing catastrophic failures. The attacker could falsify readings, change settings, disrupt communication protocols, and tamper with a host of other data as well. In their method, Kafka, Apache Spark, and R play a significant role in data accumulation and data analysis and make sense of the observed data. The observed data can be visualized and the anomalous data can be observed with the introduction of a tri-modular network. From its collection to its visualization, the data becomes clearer and is less susceptible to attacks originating externally because of the complexity at play here. This system can complement the current cyber grid systems and build toward a far more secure system in the future, which is less susceptible to attacks.

10. Discussion and conclusion

In this paper, we have presented numerous AI-based attack scenarios. We investigated and presented the technical background underl.ying those attacks and the potential mitigation strategies to address such attacks. The cyber field is evolving at a rapid rate and it is, therefore, uncertain which future scenarios it can be used in. A campaign exists in which AI researchers work together and tend not to contribute toward research in matters of their respective specialties, which may bring unwarranted negative social impacts on their surroundings. Such steps are welcome, but they deal only with the effects of issues and not with the root cause [16].

In order to control the advancement in AI's weaponization, a few compromises must be made amongst superpowers, which include checks and balances that put national and international stability and well-being first and politics and power dynamics second. It is a difficult road to traverse as different countries do not allow themselves to be imitated by others. Nevertheless, this needs to be done for the security and sanctity and more manageable research and development of controlled AI arms infrastructure. Most AI work that is conducted nowadays is in the public eye. It is difficult to obtain consensus of the world powers to decrease their involvement in such arms race because they may or may not abide by the treaty's rules, as is evident from the SALT II treaty [117].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Burton Joe, Soare Simona R. Understanding the strategic implications of the weaponization of artificial intelligence. In: 2019 11th international conference on cyber conflict (CyCon), vol. 900. IEEE; 2019, p. 1–17.

[2] China Releases video of new barrage swarm drone launcher. 2020, https://www.forbes.com/sites/davidhambling/2020/10/14/china-releases-video-of-new-barrage-swarm-drone-launcher/#484d44822ad7. [Accessed on 10/19/2020].

[3] Neff Gina, Nagy Peter. Automation, algorithms, and politics—talking to bots: Symbiotic agency and the case of Tay. Int J Commun 2016;10:17.

[4] Price Rob. Microsoft is deleting its AI chatbot's incredibly racist tweets. Bus Insider 2016.

[5] Bessi Alessandro, Ferrara Emilio. Social bots distort the 2016 US presidential election online discussion. First Monday 2016;21(11–7).

[6] Persily Nathaniel. The 2016 US Election: Can democracy survive the internet? J Democr 2017;28(2):63–76.

[7] Yamin Muhammd Mudassar, Katt Basel, Kianpour Mazaher. Cyber weapons storage mechanisms. In: International conference on security, privacy and anonymity in computation, communication and storage. Springer; 2019, p. 354–67.

[8] Rid Thomas, McBurney Peter. Cyber-weapons. RUSI J. 2012;157(1):6–13.

[9] Mirsky Yisroel, Mahler Tom, Shelef Ilan, Elovici Yuval. CT-GAN: Malicious tampering of 3D medical imagery using deep learning. 2019, arXiv preprint arXiv:1901.03597.

[10] Talk to transformer. 2019, https://talktotransformer.com/. [Accessed on 10/28/2019].

[11] Solaiman Irene, Brundage Miles, Clark Jack, Askell Amanda, Herbert-Voss Ariel, Wu Jeff, et al. Release strategies and the social impacts of language models. 2019, arXiv preprint arXiv:1908.09203.

[12] Yamin Muhammad Mudassar, Katt Basel. Ethical problems and legal issues in development and usage autonomous adversaries in cyber domain. CEUR Workshop Proceedings; 2019.

[13] Biggio Battista, Corona Igino, Maiorca Davide, Nelson Blaine, Šrndić Nedim, Laskov Pavel, et al. Evasion attacks against machine learning at test time. In: Joint european conference on machine learning and knowledge discovery in databases. Springer; 2013, p. 387–402.

[14] The STRIDE threat model—microsoft docs. 2020, https://docs.microsoft.com/en-us/previous-versions/commerce-server/ee823878(v=cs.20)?redirectedfrom=MSDN. [Accessed on 06/03/2020].

[15] Defense advanced research projects agency. 2019, https://www.darpa.mil/. [Accessed on 11/16/2019].

[16] Geist Edward Moore. It's already too late to stop the AI arms race—We must manage it instead. Bull At Sci 2016;72(5):318–21.

[17] Guzman Carlos S, Gaffe John C. Correlation of experimental and finite element modal analysis of the phalanx M61A1 close-in weapon system. Technical report, NAVAL POSTGRADUATE SCHOOL MONTEREY CA; 1995.

[18] Tomasik Brian. International cooperation vs. AI arms race. Foundational Research Institute; 2013.

[19] Li Jian-hua. Cyber security meets artificial intelligence: a survey. Front Inf Technol Electron Eng 2018;19(12):1462–74.

[20] Duddu Vasisht. A survey of adversarial machine learning in cyber warfare. Def Sci J 2018;68(4):356–66.

[21] Abbas Naveed Naeem, Ahmed Tanveer, Shah Syed Habib Ullah, Omar Muhammad, Park Han Woo. Investigating the applications of artificial intelligence in cyber security. Scientometrics 2019;1–23.

[22] Bekerman Dmitri, Shapira Bracha, Rokach Lior, Bar Ariel. Unknown malware detection using network traffic classification. In: 2015 IEEE conference on communications and network security (CNS). IEEE; 2015, p. 134–42.

[23] Ullah Mohib, Ullah Habib, Khan Sultan Daud, Cheikh Faouzi Alaya. Stacked lstm network for human activity recognition using smartphone data. In: 2019 8th European workshop on visual information processing (EUVIP). IEEE; 2019, p. 175–80.

[24] Ullah Habib, Altamimi Ahmed B, Uzair Muhammad, Ullah Mohib. Anomalous entities detection and localization in pedestrian flows. Neurocomputing 2018;290:74–86.

[25] Ullah Habib. Crowd motion analysis: Segmentation, anomaly detection, and behavior classification [Ph.D. thesis], University of Trento; 2015.

[26] Abdallah Aisha, Maarof Mohd Aizaini, Zainal Anazida. Fraud detection system: A survey. J Netw Comput Appl 2016;68:90–113.

[27] Khan Wilayat, Ullah Habib, Ahmad Aakash, Sultan Khalid, Alzahrani Abdullah J, Khan Sultan Daud, et al. Crashsafe: a formal model for proving crash-safety of android applications. Hum-Centr Comput Inf Sci 2018;8(1):21.

[28] Biggio Battista, Roli Fabio. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognit 2018;84:317–31.

[29] Ullah Mohib, Mohammed Ahmed Kedir, Cheikh Faouzi Alaya, Wang Zhaohui. A hierarchical feature model for multi-target tracking. In: 2017 IEEE international conference on image processing (ICIP). IEEE; 2017, p. 2612–6.

[30] Yi Xin, Walia Ekta, Babyn Paul. Generative adversarial network in medical imaging: A review. Med Image Anal 2019;101552.

[31] He Tong, Zhang Zhi, Zhang Hang, Zhang Zhongyue, Xie Junyuan, Li Mu. Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2019. p. 558–67.

[32] Khan Sultan Daud, Ullah Habib, Uzair Mohammad, Ullah Mohib, Ullah Rehan, Cheikh Faouzi Alaya. Disam: Density independent and scale aware model for crowd counting and localization. In: 2019 IEEE International conference on image processing (ICIP). IEEE; 2019, p. 4474–8.

[33] Sun Yu, Wohlberg Brendt, Kamilov Ulugbek S. An online plug-and-play algorithm for regularized image reconstruction. IEEE Trans Comput Imaging 2019;5(3):395–408.

[34] Khan Sultan Daud, Ullah Habib. A survey of advances in vision-based vehicle re-identification. Comput Vis Image Underst 2019;182:50–63.

[35] Khan Sultan Daud, Ullah Habib, Ullah Mohib, Conci Nicola, Cheikh Faouzi Alaya, Beghdadi Azeddine. Person head detection based deep model for people counting in sports videos. In: 2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE; 2019, p. 1–8.

[36] Ullah Mohib, Mohammed Ahmed, Alaya Cheikh Faouzi. Pednet: A spatio-temporal deep convolutional neural network for pedestrian segmentation. J Imaging 2018;4(9):107.

[37] Ben-Cohen Avi, Klang Eyal, Raskin Stephen P, Soffer Shelly, Ben-Haim Simona, Konen Eli, et al. Cross-modality synthesis from CT to PET using FCN and GAN networks for improved automated lesion detection. Eng Appl Artif Intell 2019;78:186–94.

[38] Goodfellow Ian, Pouget-Abadie Jean, Mirza Mehdi, Xu Bing, Warde-Farley David, Ozair Sherjil, et al. Generative adversarial nets. In: Advances in neural information processing systems. 2014, p. 2672–80.

[39] Ullah Habib, Khan Sultan Daud, Ullah Mohib, Mahmud Maqsood, Cheikh Faouzi Alaya. Generative adversarial networks: A short review. 2020.

[40] Nash John F, et al. Equilibrium points in n-person games. Proc Natl Acad Sci 1950;36(1):48–9.

[41] Deng Jiankang, Zhou Yuxiang, Zafeiriou Stefanos. Marginal loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2017. p. 60–8.

[42] Hinton G, Srivastava Nitish, Swersky Kevin. Rmsprop: Divide the gradient by a running average of its recent magnitude. Neural Netw Mach Learn, Coursera lecture 6e 2012.

[43] Kingma Diederik P, Ba Jimmy. Adam: A method for stochastic optimization. 2014, arXiv preprint arXiv:1412.6980.

[44] Woo Sanghyun, Park Jongchan, Lee Joon-Young, So Kweon In. Cbam: Convolutional block attention module. In: Proceedings of the european conference on computer vision (ECCV); 2018. p. 3–19.

[45] Chachra Anjali, Sharma Deepak. Applications of machine learning algorithms for countermeasures to cyber attacks. 2019, Available at SSRN 3370181.

[46] Papernot Nicolas, McDaniel Patrick, Goodfellow Ian, Jha Somesh, Celik Z Berkay, Swami Ananthram. Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on asia conference on computer and communications security. ACM; 2017, p. 506–19.

[47] Finlayson Samuel G, Chung Hyung Won, Kohane Isaac S, Beam Andrew L. Adversarial attacks against medical deep learning systems. 2018, arXiv preprint arXiv:1804.05296.

[48] Bose Avishek Joey, Aarabi Parham. Adversarial attacks on face detectors using neural net based constrained optimization. In: 2018 IEEE 20th international workshop on multimedia signal processing (MMSP). IEEE; 2018, p. 1–6.

[49] Christakopoulou Konstantina, Banerjee Arindam. Adversarial attacks on an oblivious recommender. In: Proceedings of the 13th ACM conference on recommender systems; 2019. p. 322–30.

[50] Li Juncheng B, Qu Shuhui, Li Xinjian, Kolter J Zico, Metze Florian. Adversarial music: Real world audio adversary against wake-word detection system. 2019, arXiv preprint arXiv:1911.00126.

[51] Piplai Aritran, Chukkapalli Sai Sree Laya, Joshi Anupam. Nattack! adversarial attacks to bypass a GAN based classifier trained to detect network intrusion. 2020, arXiv preprint arXiv:2002.08527.

[52] Rege Manjeet, Mbah Raymond Blanch K. Machine learning for cyber defense and attack. Data Analytics 2018 2018;83.

[53] Mohurle Savita, Patil Manisha. A brief study of wannacry threat: Ransomware attack 2017. Int J Adv Res Comput Sci 2017;8(5).

[54] Shokri Reza, Stronati Marco, Song Congzheng, Shmatikov Vitaly. Membership inference attacks against machine learning models. In: 2017 IEEE symposium on security and privacy (SP). IEEE; 2017, p. 3–18.

[55] Ribeiro Mauro, Grolinger Katarina, Capretz Miriam AM. Mlaas: Machine learning as a service. In: 2015 IEEE 14th international conference on machine learning and applications (ICMLA). IEEE; 2015, p. 896–902.

[56] Cloud prediction API is deprecated — prediction API (deprecated) — google cloud. 2019, https://cloud.google.com/prediction/. [Accessed on 11/16/2019].

[57] Machine learning on AWS. 2019, https://aws.amazon.com/machine-learning/. [Accessed on 11/16/2019].

[58] Zhang Yizhe, Gan Zhe, Carin Lawrence. Generating text via adversarial training. In: NIPS Workshop on adversarial training, vol. 21. 2016.

[59] Chen Liqun, Dai Shuyang, Tao Chenyang, Zhang Haichao, Gan Zhe, Shen Dinghan, et al. Adversarial text generation via feature-mover's distance. In: Advances in neural information processing systems. 2018, p. 4666–77.

[60] Yang Li-Chia, Chou Szu-Yu, Yang Yi-Hsuan. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. 2017, arXiv preprint arXiv:1703.10847.

[61] Antipov Grigory, Baccouche Moez, Dugelay Jean-Luc. Face aging with conditional generative adversarial networks. In: 2017 IEEE international conference on image processing (ICIP). IEEE; 2017, p. 2089–93.

[62] Bao Jianmin, Chen Dong, Wen Fang, Li Houqiang, Hua Gang. CVAE-GAN: fine-grained image generation through asymmetric training. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2745–54.

[63] Alzantot Moustafa, Chakraborty Supriyo, Srivastava Mani. Sensegen: A deep learning architecture for synthetic sensor data generation. In: 2017 IEEE International conference on pervasive computing and communications workshops (PerCom Workshops). IEEE; 2017, p. 188–93.

[64] Han Changhee, Hayashi Hideaki, Rundo Leonardo, Araki Ryosuke, Shimoda Wataru, Muramatsu Shinichi, et al. GAN-based synthetic brain MR image generation. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE; 2018, p. 734–8.

[65] Frid-Adar Maayan, Diamant Idit, Klang Eyal, Amitai Michal, Goldberger Jacob, Greenspan Hayit. GAN-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. Neurocomputing 2018;321:321–31.

[66] Ouyang Xi, Cheng Yu, Jiang Yifan, Li Chun-Liang, Zhou Pan. Pedestrian-synthesis-gan: Generating pedestrian data in real scene and beyond. 2018, arXiv preprint arXiv:1804.02047.

[67] Barsoum Emad, Kender John, Liu Zicheng. HP-GAN: Probabilistic 3D human motion prediction via GAN. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2018. p. 1418–27.

[68] Li Yitong, Min Martin Renqiang, Shen Dinghan, Carlson David E, Carin Lawrence. Video generation from text. In: AAAI, vol. 2. 2018, p. 5.

[69] Donahue Chris, McAuley Julian, Puckette Miller. Adversarial audio synthesis. 2018, arXiv preprint arXiv:1802.04208.

[70] Ghorbani Amirata, Natarajan Vivek, Coz David, Liu Yuan. DermGAN: synthetic generation of clinical skin images with pathology. In: Machine learning for health workshop. PMLR; 2020, p. 155–70.

[71] Ying Xingde, Guo Heng, Ma Kai, Wu Jian, Weng Zhengxin, Zheng Yefeng. X2CT-GAN: reconstructing CT from biplanar X-rays with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2019, p. 10619–28.

[72] Boutros Fadi, Damer Naser, Raja Kiran, Ramachandra Raghavendra, Kirchbuchner Florian, Kuijper Arjan. Iris and periocular biometrics for head mounted displays: Segmentation, recognition, and synthetic data generation. Image Vis Comput 2020;104:104007.

[73] BishopFox/deephack: PoC code from DEF CON 25 presentation. 2020, https://github.com/BishopFox/deephack. [Accessed on 10/19/2020].

[74] CyberWarefare/DeepLocker: DeepLocker - Deep learning based malware. 2020, https://github.com/CyberWarefare/DeepLocker. [Accessed on 10/19/2020].

[75] gyoisamurai/GyoiThon: GyoiThon is a growing penetration test tool using Machine Learning. 2020, https://github.com/gyoisamurai/GyoiThon. [Accessed on 10/19/2020].

[76] ThoughtfulDev/EagleEye: Stalk your friends. Find their Instagram, FB and Twitter Profiles using Image Recognition and Reverse Image Search. 2020, https://github.com/ThoughtfulDev/EagleEye. [Accessed on 10/19/2020].

[77] yanminglai/Malware-GAN: Realization of paper: "Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN" 2017. 2020, https://github.com/yanminglai/Malware-GAN. [Accessed on 10/19/2020].

[78] mindcrypt/uriDeep: Unicode encoding attacks with machine learning. 2020, https://github.com/mindcrypt/uriDeep. [Accessed on 10/19/2020].

[79] machine_learning_security/DeepExploit at master . 13o-bbr-bbq/machine_learning_security. 2020, https://github.com/13o-bbr-bbq/machine_learning_security/tree/master/DeepExploit. [Accessed on 10/19/2020].

[80] machine_learning_security/Generator at master . 13o-bbr-bbq/machine_learning_security. 2020, https://github.com/13o-bbr-bbq/machine_learning_security/tree/master/Generator. [Accessed on 10/19/2020].

[81] Griffin Slade E, Rackley Casey C. Vishing. In: Proceedings of the 5th annual conference on information security curriculum development. ACM; 2008, p. 33–5.

[82] Text to speech and AI powered deepfakes - towards data science. 2019, https://towardsdatascience.com/text-to-speech-and-ai-powered-deepfakes-9f83bac207d6. [Accessed on 11/16/2019].

[83] Parkin Simon. The rise of the deepfake and the threat to democracy. Guardian 2019;22.

[84] Lyrebird: Ultra-realistic voice cloning and text to speech — Descript. 2019, https://www.descript.com/lyrebird-ai?source=lyrebird. [Accessed on 11/16/2019].

[85] Einstein platform services. 2020, https://metamind.readme.io/. [Accessed on 10/19/2020].
[86] A GPT-3 bot posted comments on reddit for a week and no one noticed — MIT technology review. 2020, https://www.technologyreview.com/2020/10/08/1009845/a-gpt-3-bot-posted-comments-on-reddit-for-a-week-and-no-one-noticed/. [Accessed on 10/19/2020].
[87] High Rob. The era of cognitive systems: An inside look at IBM watson and how it works. IBM Corporation, Redbooks 2012.
[88] AlphaGo — DeepMind. 2020, https://deepmind.com/research/case-studies/alphago-the-story-so-far. [Accessed on 10/19/2020].
[89] Dasgupta Dipankar. AI Vs AI: Viewpoints. 2019.
[90] Musk Elon, et al. An integrated brain-machine interface platform with thousands of channels. J Med Internet Res 2019;21(10):e16194.
[91] Luo Suyuan, Lin Xudong, Zheng Zunxin. A novel CNN-DDPG based AI-trader: Performance and roles in business operations. Transp Res E 2019;131:68–79.
[92] Scherhag Ulrich, Rathgeb Christian, Busch Christoph. Towards detection of morphed face images in electronic travel documents. In: 2018 13th IAPR international workshop on document analysis systems (DAS). IEEE; 2018, p. 187–92.
[93] Pentagon to pit AI against human pilots in live fighter trials. 2020, https://www.defensenews.com/artificial-intelligence/2020/09/09/dod-to-pit-ai-vs-human-pilots-in-live-fighter-trials-by-2024/. [Accessed on 10/19/2020].
[94] Avgerinos Thanassis, Brumley David, Davis John, Goulden Ryan, Nighswander Tyler, Rebert Alex, et al. The mayhem cyber reasoning system. IEEE Secur Privacy 2018;16(2):52–60.
[95] Ali Sher, Ahmad Ayaz, Faheem Yasir, Altaf Muhammad, Ullah Habib. Energy-efficient RRH-association and resource allocation in D2D enabled multi-tier 5G C-RAN. Telecommun Syst 2019;1–15.
[96] All Info - H.R.2231 - 116th Congress (2019-2020): Algorithmic Accountability Act of 2019 — Congress.gov — Library of Congress. 2019, https://www.congress.gov/bill/116th-congress/house-bill/2231/all-info. [Accessed on 11/16/2019].
[97] Marion Nancy E. The council of Europe's cyber crime treaty: An exercise in symbolic legislation. Int J Cyber Criminol 2010;4(1/2):699.
[98] Chatila Raja, Firth-Butterflied Kay, Havens John C, Karachalios Konstantinos. The IEEE global initiative for ethical considerations in artificial intelligence and autonomous systems [standards]. IEEE Robot Autom Mag 2017;24(1):110.
[99] Falco Gregory, Viswanathan Arun, Caldera Carlos, Shrobe Howard. A master attack methodology for an ai-based automated attack planner for smart cities. IEEE Access 2018;6:48360–73.
[100] Yamin Muhammad Mudassar, Katt Basel, Torseth Espen, Gkioulos Vasileios, Kowalski Stewart James. Make it and break it: An IoT smart home testbed case study. In: Proceedings of the 2nd international symposium on computer science and intelligent control; 2018. p. 1–6.
[101] Ullah Habib, Ullah Mohib, Afridi Hina, Conci Nicola, De Natale Francesco GB. Traffic accident detection through a hydrodynamic lens. In: 2015 IEEE international conference on image processing (ICIP). IEEE; 2015, p. 2470–4.
[102] Polson Nicholas G, Sokolov Vadim O. Deep learning for short-term traffic flow prediction. Transp Res C 2017;79:1–17.
[103] Ullah Habib, Khan Sultan Daud, Ullah Mohib, Cheikh Faouzi Alaya, Uzair Muhammad. Two stream model for crowd video classification. In: 2019 8th European workshop on visual information processing (EUVIP). IEEE; 2019, p. 93–8.
[104] Xie Kefan, Mei Yanlan, Gui Ping, Liu Yang. Early-warning analysis of crowd stampede in metro station commercial area based on internet of things. Multimedia Tools Appl 2019;78(21):30141–57.
[105] Ullah Mohib, Alaya Cheikh Faouzi. A directed sparse graphical model for multi-target tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2018. p. 1816–23.
[106] Ullah Mohib, Kedir Mohammed Ahmed, Cheikh Faouzi Alaya. Hand-crafted vs deep features: A quantitative study of pedestrian appearance model. In: 2018 colour and visual computing symposium (CVCS). IEEE; 2018, p. 1–6.
[107] Miglani Arzoo, Kumar Neeraj. Deep learning models for traffic flow prediction in autonomous vehicles: A review, solutions, and challenges. Veh Commun 2019;20:100184.
[108] Ullah Habib, Ullah Mohib, Conci Nicola. Real-time anomaly detection in dense crowded scenes. In: Video surveillance and transportation imaging applications 2014. 9026, International Society for Optics and Photonics; 2014, 902608.
[109] Yamin Muhammad Mudassar, Shalaginov Andrii, Katt Basel. Smart policing for a smart world opportunities, challenges and way forward. In: Future of information and communication conference. Springer; 2020, p. 532–49.
[110] Yamin Muhammad Mudassar, Katt Basel, Sattar Kashif, Ahmad Maaz Bin. Implementation of insider threat detection system using honeypot based sensors and threat analytics. In: Future of information and communication conference. Springer; 2019, p. 801–29.
[111] deep-pwning - Metasploit for Machine Learning. 2020, https://www.kitploit.com/2016/11/deep-pwning-metasploit-for-machine.html. [Accessed on 10/19/2020].
[112] Yamin Muhammad Mudassar, Katt Basel, Gkioulos Vasileios. Cyber ranges and security testbeds: Scenarios, functions, tools and architecture. Comput Secur 2020;88:101636.
[113] Chen Yu-Hung, Chen Jiann-Liang. AI@ ntiPhish—Machine learning mechanisms for cyber-phishing attack. IEICE Trans Inf Syst 2019;102(5):878–87.
[114] Chomiak-Orsa Iwona, Rot Artur, Blaicke Bartosz. Artificial intelligence in cybersecurity: The use of AI along the cyber kill chain. In: International conference on computational collective intelligence. Springer; 2019, p. 406–16.
[115] Yadav Tarun, Rao Arvind Mallari. Technical aspects of cyber kill chain. In: International symposium on security in computing and communication. Springer; 2015, p. 438–52.
[116] Sundararajan Aditya, Khan Tanwir, Aburub Haneen, Sarwat Arif I, Rahman Shahinur. A tri-modular human-on-the-loop framework for intelligent smart grid cyber-attack visualization. In: SoutheastCon 2018. IEEE; 2018, p. 1–8.
[117] Brito Dagobert L, Intriligator Michael D. Strategic arms limitation treaties and innovations in weapons technology. Publ Choice 1981;37(1):41–59.

Muhammad Mudassar Yamin is currently doing his Ph.D. at the Department of Information and Communication Technology at the Norwegian University of Science and Technology. He is a member of the system security research group and the focus of his research is system security, penetration testing, security assessment, intrusion detection. Before joining NTNU, Mudassar was an information security consultant and served multiple government and private clients. He holds multiple cyber security certifications like OSCP, LPT-MASTER, CEH, CHFI, CPTE, CISSO, CBP.

Mohib Ullah received the bachelor's degree in electronic and computer engineering from the Politecnico di Torino, Italy, in 2012, the master's degree in telecommunication engineering from the University of Trento, Italy, in 2015, and the Ph.D. degree in computer science from the Norwegian University of Science and Technology (NTNU), Norway, in 2019. He is currently a Postdoctoral Research Fellow with NTNU, where he is involved in several industrial projects related to video surveillance. He teaches courses on machine learning and computer vision. His research interests include medical imaging, video surveillance, especially, crowd analysis, object segmentation, behavior classification, and tracking. In these research areas, he published more than 30 peer-reviewed journals, conferences, and workshop articles. He served as a Program Committee Member for the International Workshop on Computer Vision in Sports (CVsports). He also served as a Chair for the Technical Program at the European Workshop on Visual Information Processing. He is the Reviewer of well-reputed conferences and journals (Neurocomputing (Elsevier), Neural Computing and Applications (Elsevier), Multimedia Tools and Applications (Spring), IEEE ACCESS, the Journal of Imaging, Sensors, IEEE CVPRw, IEEE ICIP, and IEEE AVSS).

Habib Ullah received the Ph.D. degree from the University of Trento, Italy, in 2015, and the M.Sc. degree in computer engineering from Hanyang University, Seoul, South Korea, in 2009. He is currently working as an Assistant Professor with the University of Ha'il, Saudi Arabia. His research interests include computer vision and machine learning.

Basel Katt is currently working as an Associate Professor at the Department of Information and Communication Technology at the Norwegian University of Science and Technology. He is the technical project leader of Norwegian cyber range. Focus of his research areas are software security and security testing, software vulnerability analysis model driven software development and model driven security, access control, usage control and privacy protection, security monitoring, policies, languages, models and enforcement.