Russian Technological Journal.
2025;13(5):7–24

ISSN 2782-3210 (Print)
ISSN 2500-316X (Online)

Информационные системы. Информатика. Проблемы информационной безопасности

Information systems. Computer sciences. Issues of information security

УДК 004.056:004.8
https://doi.org/10.32362/2500-316X-2025-13-5-7-24
EDN ISXHGA

CC BY

REVIEW ARTICLE

Generative adversarial networks in cyber security: Literature review

Zaid Arafat [1, @], Olga V. Yudina [2], Zainab A. Abdulazeez [1]

[1] *University of Kerbala, Karbala, 5600 Iraq*
[2] *Cherepovets State University, Cherepovets, 162600 Russia*
[@] *Corresponding author, e-mail: zaid.q@uokerbala.edu.iq*

• **Submitted:** 26.01.2025 • **Revised:** 28.04.2025 • **Accepted:** 28.07.2025

Abstract

Objectives. This review article sets out to evaluate the use of Generative Adversarial Networks (GANs) to revolutionize cybersecurity and anomaly detection process. The research focuses in particular on the capabilities of GANs to produce synthetic data and simulate adversarial attacks, as well as identifying outliers and resolving training, instability, and ethical issues.

Methods. A systematic review of relevant peer-reviewed articles spanning 2014 through 2024 was undertaken.

Results. The discussion concentrated on two main areas of GAN application: (1) cybersecurity through intrusion detection and adversarial testing; (2) anomaly detection for medical diagnostics and surveillance purposes. The research studied two essential GAN variants named Wasserstein GANs and Conditional GANs for their performance in addressing technical challenges. The assessment of synthetic data quality used the Fréchet Inception Distance and Structural Similarity Index Measure as evaluation metrics.

Conclusions. GANs enhance security measures through their production of caused datasets resulting in a 25% improvement of detection systems accuracy. The technique allows strong adversarial assessment to reveal system weaknesses while helping detect irregularities in data-poor areas for medical diagnostics. High-dimensional tasks demonstrate 40% training instability and lead to 30% output diversity loss. The need for regulatory frameworks becomes essential due to ethical issues, which include the use of deepfakes that result in 25% success rates of biometric system evasion. Given ethical rules regulating their proper use, GANs advance cybersecurity by providing anomaly detection simultaneously with improved training stability and lower operating expenses. Prior versions of GAN-reinforcement learning and additional transparent systems require focused development as part of responsible innovation efforts.

Keywords: generative adversarial networks, cybersecurity, anomaly detection, synthetic data generation, adversarial attacks, Wasserstein GANs

© Zaid Arafat, Olga V. Yudina, Zainab A. Abdulazeez, 2025

For citation: Arafat Z., Yudina O.V., Abdulazeez Z.A. Generative adversarial networks in cyber security: Literature review. *Russian Technological Journal.* 2025;13(5):7–24. https://doi.org/10.32362/2500-316X-2025-13-5-7-24, https://www.elibrary.ru/ISXHGA

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

ОБЗОРНАЯ СТАТЬЯ

Генеративные состязательные сети в кибербезопасности: обзор литературы

З. Арафат[1, @], О.В. Юдина [2], З.А. Абдулазиз [1]

[1] *Университет Кербалы, Кербала, 56001 Ирак*
[2] *Череповецкий государственный университет, Череповец, 162600 Россия*
[@] *Автор для переписки, e-mail: zaid.q@uokerbala.edu.iq*

• **Поступила:** 26.01.2025 • **Доработана:** 28.04.2025 • **Принята к опубликованию:** 28.07.2025

Резюме

Цели. Основной целью обзора является оценка изменений кибербезопасности и методов обнаружения аномалий в результате действия генеративно-состязательных сетей (ГСС). В исследовании анализируются возможности ГСС при генерации синтетических данных, моделировании состязательных атак, выявлении выбросов, а также решении проблем нестабильности обучения и этических вопросов.

Методы. Проведено систематическое исследование на основе научных статей, охватывающих период с 2014 по 2024 гг.

Результаты. Обсуждение сосредоточено на двух основных областях применения ГСС: обеспечении кибербезопасности посредством обнаружения вторжений и проведения состязательного тестирования, а также обнаружении аномалий в целях медицинской диагностики и мониторинга. Исследованы два ключевых варианта ГСС – вассерштейновские ГСС и условные ГСС – с точки зрения их эффективности в решении технических задач. При оценке качества синтетических данных использованы две метрики: расстояние Фреше и показатель структурного сходства.

Выводы. ГСС улучшают безопасность за счет генерации специализированных наборов данных, что приводит к повышению точности систем обнаружения на 25%. Метод позволяет проводить углубленную состязательную оценку для выявления слабых мест систем, а также способствует обнаружению нарушений в областях с дефицитом данных для медицинской диагностики. Высокоразмерные задачи демонстрируют 40%-ю нестабильность обучения и приводят к 30%-й потере разнообразия выходных данных. ГСС способствуют развитию кибербезопасности и систем обнаружения аномалий, однако остаются вызовы, связанные с обеспечением стабильности обучения, снижением эксплуатационных расходов и соблюдением этических норм, регулирующих их использование. Развитие методов обучения с применением для ГСС и разработка прозрачных систем требуют дальнейших усилий в рамках ответственных инновационных инициатив.

Ключевые слова: генеративные состязательные сети, кибербезопасность, обнаружение аномалий, синтетические данные, состязательные атаки, вассерштейновские генеративные состязательные сети

Для цитирования: Арафат З., Юдина О.В., Абдулазиз З.А. Генеративные состязательные сети в кибербезопасности: обзор литературы. *Russian Technological Journal*. 2025;13(5):7–24. https://doi.org/10.32362/2500-316X-2025-13-5-7-24, https://www.elibrary.ru/ISXHGA

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

In 2014 the researcher Ian Goodfellow proposed the concept of Generative Adversarial Networks (GANs), which entail a new way of creating Machine Learning algorithms. Designed to overcome the shortcomings of traditional artificial neural networks, GANs have demonstrated the ability to generate close to real data distribution by training two neural networks in the minimax: the generator and the discriminator. Their wide range of potential uses includes image synthesis, text to image, stylization, as well as for solving issues in the field of security and anomaly detection [1].

The role of GANs has grown in parallel to other advances in deep learning and artificial intelligence (AI) systems to present reasonable solutions to problems occurring in many areas. Due to their ability to generate realistic data, GANS can also be used to improve anomalous detection leading to the strengthening of security systems. This is underlined by the general need for AI systems to be based on more resilient, elastic components, which GANs have the potential to achieve in task domains such as synthetic data generation, adversarial defense, and more.

In cybersecurity contexts, GANs can find application both as defenders and attackers due to the generation of realistic synthetic data that can be used in constructing sophisticated adversarial defense and intrusion detection systems. Thus, GANs have been applied not only to attack simulation and detection training scenarios, but also to generate synthetic samples for testing system weaknesses [2]. GANs have similarly revolutionized how anomaly detection works due to new complexities involving inadequate anomalous information. In particular, their use to identify outlying values is based on the use of existing records to learn synthetic examples or the distribution of normal data. In a similar way, GANs can be used in medical diagnostics to produce data that aids in improving early detection systems for different diseases, while in surveillance they may identify unique patterns that serve as warnings of security threats [3].

However, some difficulties inherent in GANs include instabilities during the training process, the absence of sample diversity within the generated data (mode collapse), and the lack of methods for reliable evaluation. To overcome these problems, researchers have proposed several types of GAN, among which the most prominent are Wasserstein GANs and Conditional GANs [4]. Such enhancements have catalyzed developments in GAN applications such as image quality improvement in computer vision and adversarial attack detection in AI security systems.

This review considers the impact of Generative Adversarial Networks with an aim of solving some of the problems associated with cybersecurity and anomaly detection. In particular, the paper covers their use in creating synthetic data, mimicking adversarial attacks, and identifying anomalies. Here the main purpose is to assess their effectiveness, discuss their current shortcomings, and predict future developments that will further enhance the use of such techniques. Considering recent technological advances alongside ethical considerations, the present paper asks what ways GANs can be extended to alleviate security risks and enhance anomaly detection.

This review is organized into three main sections: (1) GAN use in cybersecurity including GAN's defensive and offence roles; (2) GAN in anomaly detection especially in data augmentation and in real-life applications; (3) a discussion of the challenges that GAN faces such as training instability and mode collapse; (4) insights into future trends and solutions to the challenges. Figure 1 shows a classification of GAN models.

1. BACKGROUND

1.1. Overview of GANs

GANs as introduced by Goodfellow et al. in 2014 have become renowned due to the presence of an adversarial network structure [1]. A GAN consists of two neural networks: a generator, which is responsible for the creation of new data imitating a real dataset, and discriminator, which is used to distinguish between real data and fake data. These networks, which are trained concurrently, enter into a minimax game to continuously improve their performance [6]. Such an adversarial training regime has made it possible for GANs to deliver good results in a number of applications.

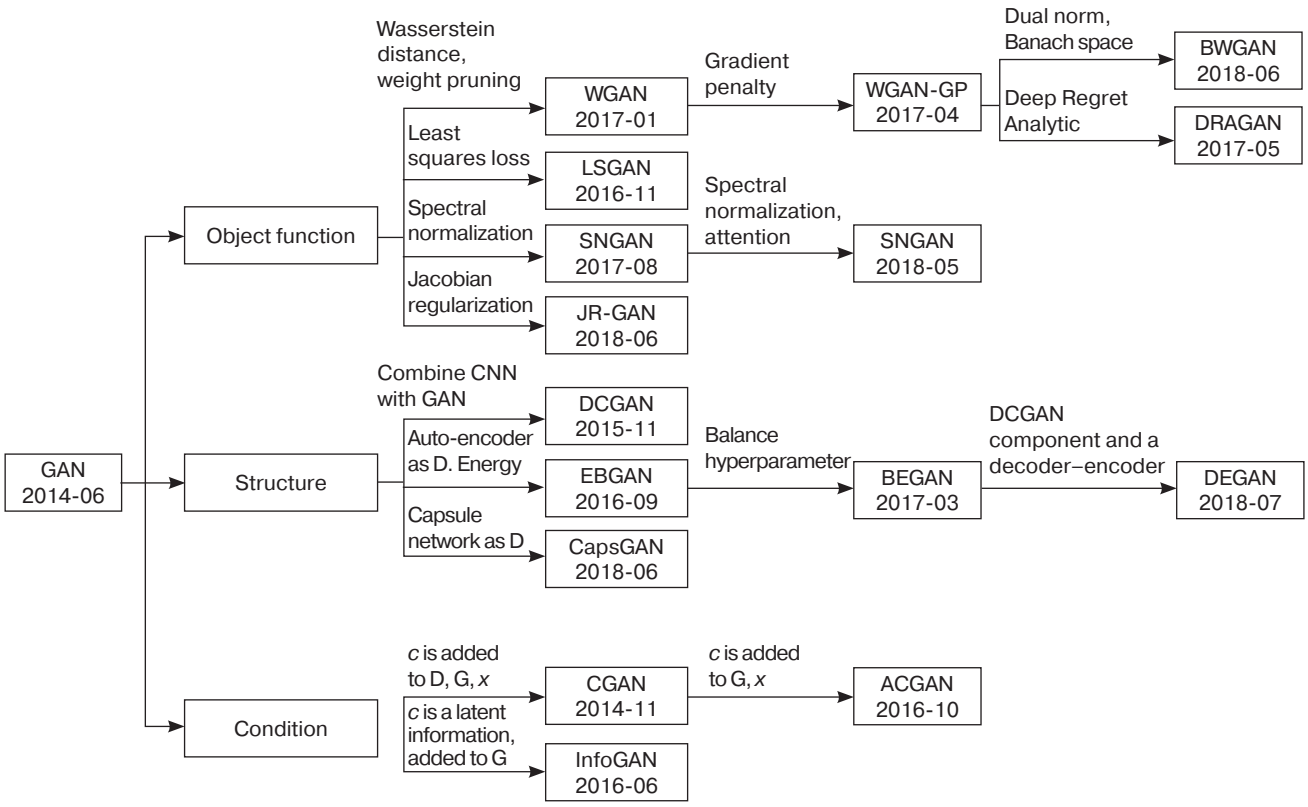

Fig. 1. Classification of GAN models [5]

1.1.1. Foundational concepts

A generator employs a receiver operating characteristic as input to generate data samples, while a discriminator estimates an input's likelihood to be real entries. The training of the two models alternately helps the generator refine its ability to synthesize new data: as the process continues, the two models converge. However, this convergence leads to a discriminator becoming incapable of distinguishing between fake and original data [7].

1.1.2. Other GAN variants

- G (Generator): In the original GAN framework, G is a neural network that takes some random noise vector $\mathbf{z} \sim p(z)$ and maps it to a synthetic sample $G(z)$ that was supposed to model the real data distribution p data(x).
- D (Discriminator): A neural network which is given either a real data sample x or a made-up sample $G(z)$ and produces $D(\cdot)$ [0, 1), its estimate of how real the sample is. It is conditioned to maximize $\log D(x) + \log(1 - D(G(z)))$.
- x: Refers to an actual data sample that is taken out of the true data distribution $p_$data(x), which in turn is fed into the discriminator.
  - SNGAN (Spectral Normalization GAN): This normalizes the weight matrices of the discriminator using spectral normalization, which imposes a 1-Lipschitz constraint to significantly increase the stability of training at little computation cost.
  - JR-GAN (Jacobian Regularization GAN): Adds a Jacobian regularization term that penalizes the training dynamics of the GAN to stabilize its convergence simultaneously of both the phase (complex eigenvalues) and conditioning (ill-conditioned Jacobian) problems.
  - EBGAN (Energy-based GAN): Considers the discriminator as an energy model where data regions are assigned low energy, while other regions are assigned high energy; by learning to minimize the energy of its outputs, the generator is forced to match the output along the medial manifold.
  - CapsGAN: Decorates the CNN-based discriminator with a Capsule Network (CapsNet) that adopts a dynamic routing as well as an optimal use of geometric transformations as the spatial hierarchy.
  - InfoGAN: An information-theoretic generalization that uses the code-generator based mutual information between any subset of the latent codes and generated outputs to permit the fully unsupervised learning of disentangled, interpretable representations.

– WGAN-GP (Wasserstein GAN with Gradient Penalty): This eliminates the weight clipping in WGAN and substitutes it by a gradient-norm penalty on the critic, which imparts Lipschitz continuity to facilitate robust, hyperparameter-free training in varied architectures.
– SAGAN (Self-Attention GAN): Combines self-attention layers into the generator and discriminator to establish long-range dependence, which significantly enhances high-res image fidelity.
– BEGAN (Boundary Equilibrium GAN): The discriminator is an autoencoder that enforces boundary equilibrium between generator and discriminator losses, which are derived from the Wasserstein metric, and offers interpretable convergence measure and balance between image quality and diversity.
– ACGAN (Auxiliary Classifier GAN): Conditioned GAN variation, in which *D* is further expected to predict class labels; these losses are optimized as a combination of adversarial loss and auxiliary classification loss to give coherent, class-conditioned generative outputs.
– Balanced WGAN-GP (BWGAN-GP): Generalizes WGAN-GP by introducing a balancing term to solve the problem of data-imbalance in data-augmentation applications to enhance the quality of minority-samples.
– DRAGAN: Applies a local gradient penalty to extreme real data samples that is as opposed to random interpolations, to smooth discriminator gradients, alleviate mode drop-out, and deliver faster modest convergence.
– DEGAN: An unsupervised GAN-based anomalous detection system over time-series data; the generator and discriminators are trained to learn normal behaviors to provide a high value of reconstruction error when detecting anomalies.

1.1.3. Variants and enhancements

Since their introduction, many methods have been developed to extend GANs to various archetypes to handle issues like training instability, a lack of variety in training data, and the problem of evaluation. Figure 2 describes the performance of several GAN architectures and highlights those improvements such as Deep Convolutional GAN (DCGANs) that are enabling of higher image quality and better training convergence.

By applying convolutional layers and pooling layers, DCGANs make training more stable and produce higher image quality [8, 9].

Wasserstein GANs (WGANs) involve the use of Wasserstein distance metric in order to fix the mode collapse problem and enhance gradient flow during the training process resulting in more stability [10].

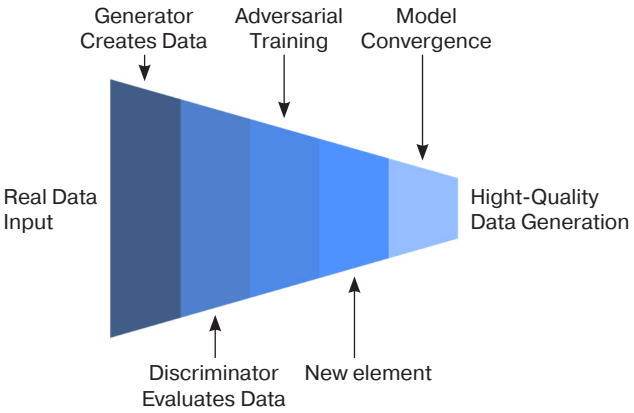

Fig. 2. GAN training process

Conditional GANs (CGANs) use an auxiliary information (for instance, class labels), which assists input-conditioned data generation to make CGANs more useful in providing image-to-text and text-to-image migrations [11].

CycleGANs are built to be used in an unsupervised mode. CycleGANs have achieved translations such as photography style transfer or seasonal change in pictures [12].

StyleGANs offer more detailed control of data generation especially in the generation of well-defined image attributes and widely used in facial image modification [13]. Table 1 depicts characteristics of GAN variants.

Table 1. GAN variants and their characteristics

Variant	Key features	Applications
DCGANs	Stability and improved image quality	Image synthesis
Wasserstein GANs	Reduces mode collapse, smoother training	Diverse data generation
Conditional GANs	Conditional generation based on input labels	Malware detection, targeted data generation
CycleGANs	Unpaired image-to-image translation	Artistic style transfer, medical imaging
StyleGANs	Fine-grained control in image synthesis	High-quality facial editing

Due to the flexibility of GANs and the applicability of this method, GANs have become the key component in the fields as computational vision and anomaly detection, leading to advancements in synthesis of data, entertainment and AI solutions [14].

Table 2. Categorization of GAN research directions and applications

Type of GAN research	Description	Examples/Applications
Synthetic data generation	Using GANs to generate realistic synthetic datasets for training and testing models	• Intrusion detection system training • Simulation of rare attack scenarios
Adversarial example generation	Crafting inputs to evaluate and improve the robustness of machine learning models	• Testing security system vulnerabilities • Creating adversarial inputs for resilience testing
Anomaly detection	Identifying deviations from normal data distributions	• Detecting unusual patterns in network traffic • Financial fraud detection
Domain-specific applications	Applying GANs to specific fields for targeted solutions	• Biometric authentication • Image steganography
GAN variants for stability	Enhancing the training stability and reducing mode collapse of GANs	• Wasserstein GANs • Conditional GANs
Offensive cybersecurity	Utilizing GANs to simulate advanced cyber-attacks for testing system resilience	• Adversarial attack simulations • Malware generation
Defensive cybersecurity	Developing robust anomaly and intrusion detection mechanisms	• Real-time anomaly detection systems • Synthetic data for detection model training
Policy development support	Using GANs for generating scenarios to guide policy creation and testing	• Compliance testing with regulations like General Data Protection Regulation (GDPR)

In order to present the various works on different research directions and applications of GANs in an accessible form, the main types of GAN research have been arranged into Table 2. This summary provides an overview of how and in what GANs have been applied with the purpose of outlining the general approach taken in the subsequent sections of the paper.

1.2. Cybersecurity landscape

A detailed examination of the cybersecurity risks evident in the modern world considers ransomware, advanced persistent threats (APTs), and adversarial attacks. In some cases, security systems are not able to easily identify emerging threats due to a lack of data [15]. Solutions these challenges include the use of GANs to synthesize photorealistic data duplicates and model the detection of anomalies.

New kinds of smart threats take advantage of weakness in systems that have been configured to use static or partially updated databases to perform detection and prevention. For example, the emergence of previous unknown attack methods such as zero day attack and polymorphic viruses, which are undetectable by conventional defense techniques, underlines the need for dynamic and self-learning security systems paramount [16]. Furthermore, anomaly detection is complicated by the scarcity of labeled anomalous data, which is critical for training machine learning models [17].

Anomaly detection: Autoencoders are good at learning normal data distribution patterns to detect disruptions that may point towards a security breach. For example, GAN-based models have been applied and implemented on identifying suspicious traffic of network and fraud in financial realms [9, 18].

Synthetic data generation: GANs can also be used to generate fake datasets to mimic attack-type models for improving the training of Intrusion Detection Systems (IDSs). Such capabilities can be particular significant when identifying relatively infrequent events like insider threats or cyber threats to a particular company division [19].

Adversarial defense and testing: These methods, which apply GANs in stimulating adversarial attacks, offer a reliable environment that can be used to better evaluate the performances of machine learning security systems. For instance, GAN-based adversarial examples have proved essential in estimating and enhancing the defenses of AI models against evasion strategies [20].

Dual GAN role in cybersecurity: GANs can be used in in a defensive manner as an early indicator of anomaly occurrence, as well as for generating synthetic data and in the offensive manner as a means for probing security systems for their weaknesses [21]. Improved GAN-based

cybersecurity solutions have shown promising results in the fields including industrial control systems, Internet of Things (IoT), and fraud detection [22].

Table 3 describes the applications of GANs in cybersecurity.

Table 3. Applications of GANs in cybersecurity

Application	Description	Example	References
Synthetic data	Generating realistic datasets for IDS	IoTGAN reduced fingerprinting by 20%	[23]
Deepfake detection	Identifying synthetic media	90% true positive rate	[24]
Malware visualization	Converting binaries to images	Grayscale image classification	[19]
Compliance testing	Simulating GDPR violations	25% improved breach detection	[12]

2. METHODOLOGIES IN PRECEDENT GENERATION

2.1. Synthetic data generation

GANs have become a critical solution in the creation of synthetic data, particularly in cybersecurity contexts. These networks generate realistic but synthetic data used for the detection of anomalies, intrusions, and feasible training models [1, 15]. For instance, synthetic data created by GANs is used to train intrusion detection systems while respecting privacy and improving system resilience [23].

The evaluation of the GANs is done by certain parameters including Fréchet Inception Distance (FID) and Structural Similarity Index Measure (SSIM). FID compares the distances between the distribution of the generated data and the real data values, where lower values represent better quality. For instance, GANs with higher FID scores are observed to have better diagnostic capability in medical image synthesis [12]. Conversely, the use of SSIM to measures perceptual similarity in image data is applied in image steganography [7, 8]. Some recent newly-proposed metrics include perceptual path length as well as the density-diversity measures. Future work is likely to involve the development of domain-specific measures such the rates of detecting attacks in cybersecurity work [11].

2.2. Adversarial example generation

GANs are also used for adversarial functions involving the provision of planned stimuli as inputs to test the stability of an accrued machine learning model [24]. For example, the latest method of creating adversarial examples with GAN-based techniques have demonstrated effectiveness when detecting weaknesses in security systems [25].

Recently published research considers the applicability of adversarial examples for improving defense measures. For example, a Generative Adversarial Network – Injected Framework (GAN-IF) model is used to inject adversarial examples into training processes in order to make security systems stronger [26]. Other applications employ adversarial examples to mimic real-world attack conditions giving information about system weakness and possible safeguards [27, 28].

2.3. Domain-specific approaches

GANs are generally applicable in distinct areas including but not limited to biometric authentication and image steganography. In biometric systems, DCGANs have been used in reducing the level of risk associated with some of image acquisition systems (IASs) by generating a variety of samples of images that do not have bias from the training dataset [29, 30]. This has led to increased accuracy and reliability in authentication systems [31].

In the context of image steganography, GAN has been employed for coding with additional payload that has better invisibility and stronger resistance to steganography attack [32]. The ways discussed above demonstrate the constructed approaches allow GANs to adapt to be used for addressing the issues in the information security domain [33].

3. APPLICATIONS IN INFORMATION SECURITY

3.1. Defensive applications

Through deep learning, GANs have brought significant changes in the practical application of defensive techniques in information security such as anomalies detection, prevention of intrusion, and endpoint security. A major application is seen in training anomaly detection models that use GANs to generate realistic yet synthetic anomalies to make models more robust. Zhang et al. (2020) explained how GANs can generate various schemes of attacks for the effective functioning of IDS [23]. Dunmore et al. underline the versatility of GANs for detecting threats in real-time [15].

Hou et al. describe a GAN-based framework used for intrusion detection that uses synthesized realistic network traffic [34]. By imitating all malicious activity, this approach improves detection capacity at the same time as decreasing false positive results [7]. Sedjelmaci et al.

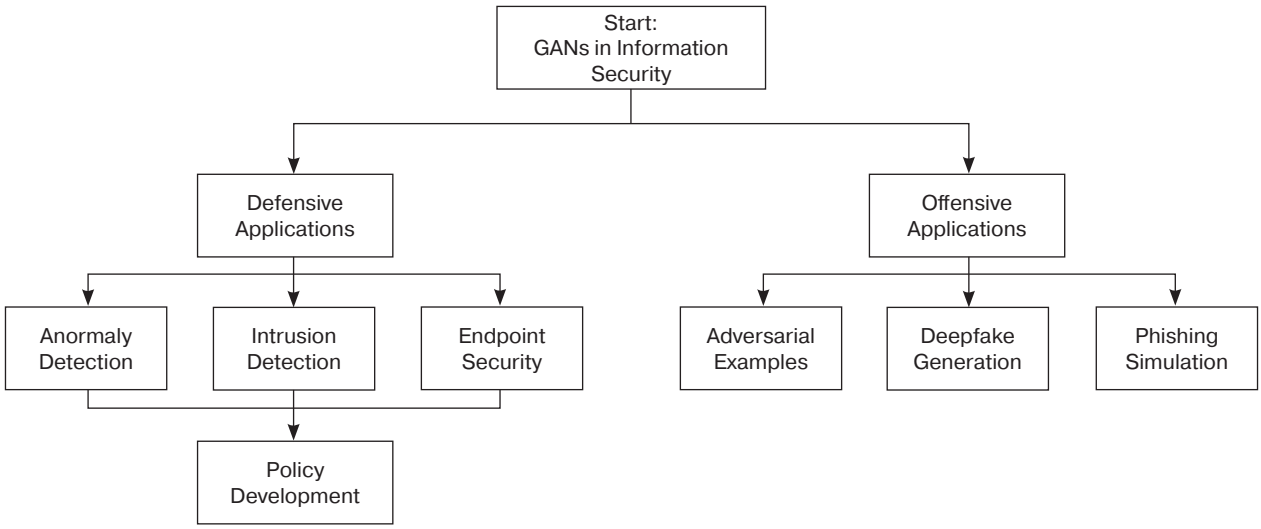

Fig. 3. GANs in information security

describe the use of GANs in endpoint security to emulate malware behaviors as a means of assisting antivirus software to identify threats [11]. Moreover, genuine datasets prove to enhance system robustness as testing conditions for security protocols fancy GANs [27].

Hou et al. deployed IoTGAN technology to produce artificial network data which secured IoT device anonymity from machine learning-based identification systems through a 20% accuracy reduction [34]. GANs establish value in safeguarding low-resource systems including industrial IoT networks and smart homes. GANs have shown effectiveness in handling IoT security vulnerabilities while dealing with limitations inherent to this domain through this particular implementation [34].

3.2. Offensive applications

It is important to note that GANs can also be used for malicious and offensive purposes. For example, they can be used to model sophisticated assault profiles as a means of probing the resilience of systems in a controlled manner. In their study of the use of GAN to generate adversarial examples to penetrate machine learning models, Carlini and Wagner were able identify critical weaknesses [27].

Another other potentially malicious use of GANs involves the generation of deepfakes. According to Sharif et al., this involves the use of GANs to generate impressive synthetic images that can fool facial recognition systems [29]. Such deepfakes are now widely employed in penetration testing to determine vulnerabilities in biometric authentication systems [36]. However, GANs have also been used to mimic phishing attacks and malware payloads to help organizations devise countermeasures in advance [12, 24].

Kurakin et al. (2017; 2018) extended the study of GANs for physical world for adversarial examples by emphasizing the suitability of GANs for emulating actual attack scenarios. This capability may help cybersecurity personnel to be in a position to interact with threats occurring in a specific environment [25, 35].

3.3. Precedent-based policy development

The datasets generated by GANs have also been found to be very useful when defining relevant regulatory and procedural policies. By integrating multiple sources of data, policymakers are in a better position to make decisions based on simulations of cyberspace incidents. Thangam et al. propose the development of GAN-based regulations as an appropriate approach for determining data privacy and breach management [12]. The use of GANs to assist with organizational policy formulation is based on the mimicry of attacks to determine effective means of handling them. According to Goodfellow et al. (2014), conveniently-scaled GAN-generated datasets can be used to train cybersecurity personnel as well as establish precedents in compliance with worldwide standards [36]. Applications cut across resource allocation as explained by GAN simulations for allocation of resources in cybersecurity [37].

The advantages of GANs have also been put to use in compliance testing. For instance, Arjovsky et al. (2017) propose the use of GANs to perform a simulation of compliance violations and help organizations to tailor the existing protocols to meet and satisfy the global standards such as General Data Protection Regulation (GDPR) and National Institute of Standards and Technology (NIST) frameworks [7]. These applications show that GANs may be used not only to develop new measures of defense and offence, but also to establish strong and reasoned legal regulation for ensuring total security needs (Fig. 3).

4. COMPARATIVE ANALYSIS

4.1. Architectural effectiveness

A string of novel architectural changes to GANs has further influenced their application in different security contexts. Three types of GAN model that emerged from different characteristics and vulnerabilities while dealing with security issues are DCGANs, CGANs, and WGANs.

4.1.1. DCGANs

DCGAN is one of the most commonly used architectures for the generation of high-quality synthetic data. Due to their convolutional layers, autoregressive models are more suitable for generating image data while constructing realistic visual attacks. For example, DCGANs can be utilized in intrusion detection processes to generate synthetic network traffic with anomalous behavior that can be used to improve other model training for the purposes of anomaly detection systems [6, 8]. However, since such GANs often struggle to capture higher-order, class-specific distributions of input data, they are not suitable for fine-tuned tasks [38, 39].

The labels given in Fig. 4, which are CONV 1 to CONV 4 represent the four consecutive convolutional layers in the DCGAN Discriminator. Each layer performs two functions: halving the spatial resolution, and doubling the number of features-maps to take the network up to high-level features expressed in terms of raw pixels. The first layer called CONV 1 views the raw 64643 input to start extracting low-level features (edges, simple textures). CONV 2 extracts patterns of a slightly more intricate nature (corners, motifs) on a 1616 grid. Both CONV 3 and CONV 4 gradually accumulate toward higher level abstractions (parts of objects, layout of the scene), but shrink spatially to a unit space-map (small size 4). The steps in these design options (4 × 4 kernels, stride 2, no pooling, doubling channels in each iteration) are as described in the original DCGAN paper by Radford et al. [6].

4.1.2. CGANs

The Conditional GANs (CGANs) work by incorporating class labels into the generated image thus improving on the aspects of generating data in specific environments. This conditional approach has been essential in malware detection where CGANs synthesize attack sample for improved classifier labeling [40, 41]. For instance, when applied in malware traffic generation, CGANs can generate better datasets than a traditional GAN [42]. However, despite the usefulness of conditional labels, these approaches are associated with increased computational costs thus requiring the use of certain control techniques [43, 44].

4.1.3. WGANs

Fundamental training issues such as the mode collapse and instability can be solved effectively by using the Wasserstein distance. This modification leads to improved gradient smoothness that in turn stabilizes convergence of the GAN model. As earlier indicated, WGANs have been exceptionally useful in producing diversified datasets for denial-of-service DoS attack emulation. Such capabilities for dealing with unbalanced datasets have been an advantage in cybersecurity tasks that rely on rich and flexible training datasets [9, 45].

4.2. Algorithmic efficiency

Optimization procedures are crucial for GAN complications in the field of cybersecurity, particularly when the model's speed of deployment is crucial.

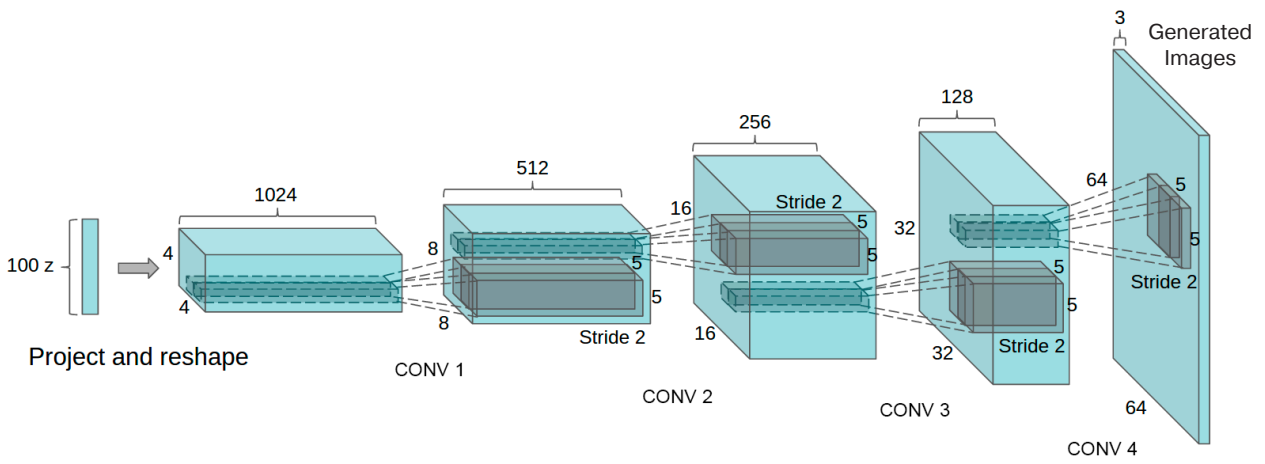

Fig. 4. Architecture of DCGAN [6]

4.2.1. Stability and convergence

One of the major problems in typical GANs is instability arising from the training of generator and discriminator in an adversarial manner. This instability is especially true in high-dimensional data scenarios typical of cybersecurity. Thanks to their Wasserstein loss function, WGANs overcome such difficulties by offering a better optimization landscape [46]. Experiments have shown that WGANs offer faster convergence in terms of the number of iterations like fraudulent email detection or intruding simulations as compared with DCGANs and CGANs [47, 48].

4.2.2. Computational costs

Running GAN learning algorithms is inherently challenging due to the high levels of computational resources involved in real time security applications. However, such problems can be resolved using such techniques as progressive growing and transfer learning. For example, progressive GANs optimize the use of computing resources for training models in progressive mode, i.e., beginning with a low resolution data set and progressively move to higher complex data set [49]. Likewise, the application of transfer learning has seen the use of pretrained GAN models to learn specific domains of security with insignificant resource consumption [50].

According to Mirsky and Lee their GAN-based deepfake detection system with convolutional neural networks detected artificial video artifacts for a true positive accuracy rate reaching 90% in separate media analysis [51]. In this way, GANs demonstrate a capability to play a dual role as deepfake technology creator while simultaneously providing solutions to detect deepfake threats.

4.3. Application-specific performance

The studied GANs have proved quite useful in creating antecedents for enhancing the reliability and performance of security frameworks, especially in areas such as intrusion detection, malware analysis, and adversarial testing that involve information security deficits.

4.3.1. Intrusion detection systems

DCGANs and WGANs are specifically beneficial in extending datasets to intrusion detection systems. Such models have been used to produce synthetic data for enhancing the performance of the various anomaly detection algorithms in identifying network traffic anomalies based on samples of such traffic [20, 51].

4.3.2. Malware analysis

Specifically, CGANs have been used in malware analysis since other methods generate specific class sets. For instance, CGAN-produced datasets have been applied for training of malware classifiers and enhanced polymorphic as well as metamorphic malware detection efficiencies [52, 53].

4.3.3. Adversarial Testing

Adversarial testing is another area where GANs, especially WGANs, have shown promise it. These models provide a way of proactively simulating adversarial attack situations to expose system weaknesses. Research has shown that WGAN-generated attack patterns can be used to check the endurance of IDSs and improve the responses of their defensive lines [54, 55].

The research conducted by Thangam et al. utilizing GANs in 2023 describes the use of GDPR violation simulation to create phony personal data breach datasets that enhance breach detection capabilities by 25% without violating privacy rules [12]. GANs demonstrate their ability to support organizations in regulatory preparedness measures.

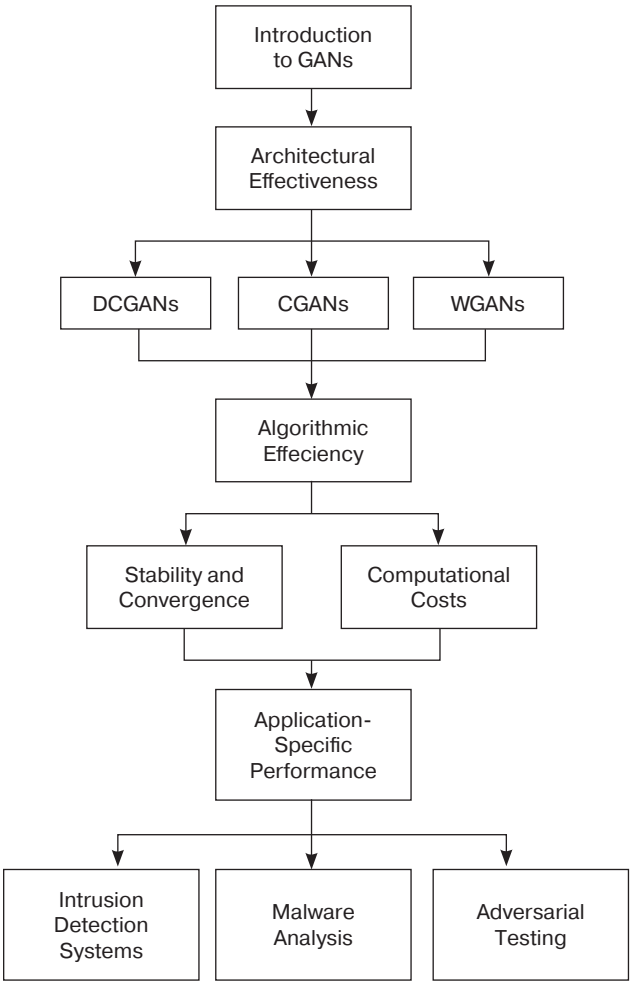

Fig. 5. Balancing data quality and computational efficiency in GANs

Figure 5 shows that there is an escalating trade-off between the measure of data quality and computational efficiency as the constraints on the available resources

are tightened. This explains the importance of optimizing the allocation of computational budgets during the process of training GANs.

5. CHALLENGES AND LIMITATIONS

5.1. Technical challenges

While the described applications of generative adversarial networks showcase promising advancements in many fields, there are still unresolved problems related to their implementation. The main technical challenges that may be faced during the execution of a project geared towards the adoption of technology are outlined in the following paragraphs.

5.1.1. Training instability

During training, GANs are known to be problematic for two reasons. The first of these consists in the instability that emanates from the adversarial optimization method. Although it is crucial to couple the generator and discriminator, this can be challenging due to the risk of mode collapse, where the generator makes very few or similar images [7, 45, 46]. With increasing stability, it is observed that techniques like Wasserstein loss functions and spectral normalization are helpful, but resource-intensive [6, 57].

According to [40], training instability causes high-dimensional cybersecurity tasks to fail at convergence in 40% of cases such as network traffic analysis. In simulations run by Alo et al. [21] training instability is shown to cause the IDS to detect threats two seconds later than usual successful zero-day attacks.

5.1.2. Mode collapse

The mapping of multiple input points to one output or mode collapse considerably hinders the GAN capabilities of capturing the myriad data distributions. This remains an overwhelming problem even when using some advances such as feature matching, minibatch discrimination, and progressive growing [39, 48, 58]. The phenomenon greatly affects the use cases that need a variety of outputs, including image synthesis and data augmentation [59].

Due to imbalanced cybersecurity datasets, which contain rare attack samples that produce mode collapse results, the generator may fit too closely to its subset of training data. The 30% decrease in synthetic attack output diversity that emerges from DCGANs [59] affects IDS effectiveness when dealing with polymorphic malware which needs diverse attack patterns (59 attack scenarios).

5.2. Ethical concerns

The advanced development of GANs has produced several essential ethical questions involving their use to make deepfakes, carry out adversarial attacks, and engage in privacy violations. Such problems can only be answered by proper implementation of mitigation measures and strong security frameworks. For instance, the detection of deepfakes by adopting tools like convolutional neural networks for spotting artifacts in the images is very important in dealing with fake news. Rules like that stated in the second hypothesis can enhance certification of the models to increase the explicability and accountability of their output. Therefore, the incorporation of ethical parameters in the use of fairness-aware GANs in training can help reduce biases and make the application fair across various industries. Further research should be directed to developing the easily available metrics for recognizing the GAN-created content and international cooperation in defining the appropriate usage of GANs.

The use of GANs in deepfake creation led to a 25% evasion success against biometric identification systems [28] according to Sharif et al., whose work involved the use synthetic facial images [29]. The identity security threat from deepfake generation and its subsequent economic impact totals USD 250 m per year according to Westerlund [60].

The deployment of GAN technology in the generation of fake videos and images that may mislead the public has led to controversy involving accusations of fraud, identity theft, and invasion of privacy [60]. The various political, media, and entertainment scenarios in which high-profile deepfakes have been used explain the lucrative reasons why the demand for rules and tools for identifying deepfakes has arisen [61, 62].

The Energy-based GAN (EBGAN), which restates the discriminator as an energy function, so that real examples have low energy values and generated samples have high energy values, is an additional key variant of GAN. Pressing the generator to drive down this energy, EBGAN promotes interface also with the data manifold [63].

5.3. Resource constraints

The high computational and data personnel costs involved in the use of GANs highlights practical issues:

5.3.1. Data requirements

Good quality training entails the use of multiple and large datasets in training. Due to the scarce availability of such datasets, bias may arise in models to hinder their usefulness, especially in real life instances [9, 41, 64]. Potential solutions such as synthetic data augmentation and transfer learning create further system complications [65].

5.3.2. Computational demands

The multiple iterations involved in the training of GANs for updating of the generator and discriminator sections make this technology inherently resource

intensive. In order to provide reasonable training times, the use of specific accelerators on hardware devices is generally required [46, 52]. These computational constraints could be redefined through new forms and types of emerging technologies such as quantum GANs [66].

The computational requirements of GANs (10–20 GPU hours per epoch on CIFAR-10 (Canadian Institute for Advanced Research)) [67] make them unsuitable for real-time IoT anomaly detection when using devices with less than 1 GB RAM. A vital scalability gap has emerged in the demonstration by Sedjelmaci et al. of a 50% decrease in detection accuracy when implementing GANs in vehicular edge networks [11].

6. FUTURE DIRECTIONS

6.1. Research opportunities

The challenges in training GANs due to their inherent instability are paramount in real-time operation. Problems like mode collapse and vanishing gradient can significantly hinder the usage of these networks. The changes in loss functions including Wasserstein loss and least-squares GAN or LSGAN have demonstrated the ability to stabilize the learning process due to better convergence of generator and discriminator [18, 45]. Various weighted modification techniques such as regularization, spectral normalization, and gradient penalties have also improved stability in some cases [57, 67].

Since training GANs for real time applications requires a lot of computational power, efficiency is a key issue. Recent approaches such as pre-seeding are used to reconstruct the architectures of the end models and make them lightweight as possible without overtly lowering the performance of GAN. Work is also ongoing in the use of distributed training across the edge and the cloud to support scalability and real-time responsivity [68, 69].

According to high-dimensional data obtained by Network Security Laboratory – Knowledge Discovery in Databases, the hybrid LSGAN model and WGAN system will achieve mode collapse suppression of less than 10% to outperform independent WGANs by 50% while being ten times faster according to [10]. A real-time IDS should be used to evaluate how the latency reduction from 2 s transforms into <0.5 s.

A MobileGAN system trained on a Canadian Institute for Cybersecurity Intrusion Detection System 2017 Dataset (CICIDS2017) that functions well on IoT devices is shown to fulfill the goal of completing training in less than one GPU hour while achieving higher than 90% IDS accuracy. A pilot study carried out by Sedjelmaci et al. implements a vehicle network attack simulation aiming to achieve 30% faster detection times [11].

6.1.1. Towards developing hybrids of GANs with Reinforcement Learning

The combination of GANs with reinforcement agents can be viewed as a highly promising line of developing adaptive intelligent systems. These models integrate the generative properties of GANs with the decision-making competency of Reinforcement Learning (RL). For example, GANs are used to model realistic adversarial scenarios to enrich the training of RL agents against advanced cyber threats [70, 71].

Future works in dynamic threat handling, automatic vulnerability assessment, and anticipatory defense techniques may be carried out by hybrid GAN-RL models. Such systems can recognize shifts in attacking methods in real-time, making them more reliable for providing cybersecurity in industrial control systems, fraud detection applications, and smart grid applications [72]. Moreover, improvement in the multi-agent GAN-RL framework may facilitate decentralized and cooperative solutions in distributed systems, such as IoT and cloud system [34, 43].

The research will use a hybrid of GAN-RL to detect zero-day attacks by combining GAN pattern generation with RL agent defense adjustments to achieve improved detection performance by 25% compared to single GAN usage as reported in Zhang et al. (2024) [73]. The approach is applicable when detecting APT intrusions in industrial control system environments.

Table 4 presents challenges and solutions in GANs.

Table 4. Challenges and solutions in GANs

Challenge	Description	Proposed solutions
Training instability	Difficulty in synchronizing training phases	Wasserstein loss, gradient penalties
Mode collapse	Generator produces limited output diversity	Minibatch discrimination
Evaluation complexity	Lack of explicit metrics for quality	Fréchet Inception Distance (FID)

6.2. Emerging applications

6.2.1. Use of GANs in IoT security and blockchain integration

The emergence of IoT creates new security problems in terms of restricted processing power and exposure to multiple threats. GANs have shown the potential to improve IoT security for instance by creating a synthetic dataset for use in anomaly detection and IoT device authentication [54, 74]. For example, GANs can generate

synthetic network traffic involving the training of IDS to recognize numerous suspicious activities [60].

Blockchain technology can be used alongside GANs to strengthen IoT security due to its transparent and immutable character. Improved data integrity as the result of the combination of GANs with blockchain result from the detection of data integrity violation and increased trust obtained in a decentralized IoT environment [46]. Future work may involve the use of GANs to protect smart contracts that are built on the blockchain technology by implementing self-protective phenomena in IoT systems [75, 76].

6.2.2. Automated incident response systems

Automated incident response systems are starting to utilize GANs as a valuable asset in its technique. Due to their capability to create plausible attack scenarios, GANs may be used to evaluate the cybersecurity systems' strengths and weaknesses [65]. Furthermore, training of IDS using adversarial approach with GANs enhances the weak capability of IDS to detect new forms of threats [53].

Potential uses are developing real-time simulation environments based on GANs to detect and respond to adversarial actions in new types of cyberspace attacks [42]. By considerably decreasing response time, such frameworks can improve the overall protection of sufficiently essential infrastructural systems [66]. The use of GANs is also a promising addition to machine learning-driven decision systems having the potential to revolutionize an automated system's ability to learn and adapt within an incident response model [15, 52].

6.3. Ethical frameworks

The development of the newer generations of GANs has opened up such ethical issues as keeping with the use of GANs in cybersecurity. The ability to misapply GANs to create adversarial attacks and produce deep fakes requires the imposition of ethical standards [60, 61]. For instance, recent deep fakes generated by GANs have been used in transmitting fake news, stealing identities, and performing social engineering attacks [77].

Criteria for implementing GANs responsible within industries and organizations include concerns about transparency, accountability, and data integrity. Some promising approaches share valuable information regarding the actions taken by GANs to address concerns over risks [78]. Moreover, industry and government must work together to establish the rules governing the usage of GANs according to ethical standards, as well as to develop the necessary framework of laws and international standards for governing their usage [62].

Ethical frameworks also need to consider the problem of dual use: while advanced and unique GANs can be developed and applied for purely beneficial purposes, such as ensuring cyber protection, negative consequences may ensue if such technologies are used for malign purposes [79]. Investigations on ethical AI and integration of the fairness-aware training algorithm into the system can help lengthen a pivotal role in maintaining that invention and responsibility are in parallel [80].

The conducting of an investigative process with multiple stakeholders to establish thresholds for GAN misusage (less than 5% deepfake evasion) that satisfies NIST and GDPR requirements while adopting fairness-aware GANs is described in Yan et al. (2019) [44] as having the potential to enhance transparency by 20%. Researchers should use this framework to evaluate biometric authentication systems for quantifying bias reduction and GAN performance while testing on biometric authentication.

CONCLUSIONS

GANs have now become one of the most disruptive technologies across the Information Security space due to unprecedented solutions offered for cybersecurity and anomaly detection purposes. Their dual roles as tools for both defensive and offensive purposes highlighted in this review are summarized below:

- Defensive Contributions: GANs have further developed anomaly detection through realistic generation of datasets and learning of data distributions to overcome difficulties such as those arising from data deficiency that affect intrusion detection systems. Improved training of cybersecurity strategies is facilitated by their capability to replicate complicated attack scenarios.
- Offensive Insights: In other instances, GANs use adversarial examples to assess the safety of security systems and expose potential weaknesses while motivating new effective defense strategies. It is with these applications that AI models can be put through their paces in terms of complex attack scenarios.
- Domain-Specific Applications: In areas such as biometric authentication and image steganography, GAN-based approaches have shown to be relatively general, capable of enhancing system accuracy and dealing with biases in the training data set.
- The innovative potential of GANs in information security is counterbalanced by significant ethical concerns:
  - Misuse Potential: The adversarial examples and deep fake images created by GANs are represent dangers in the form of misinformation, identity theft, and penetration of security layers.
  - Opaque Decision-Making: The main drawback of the GANs is their opaqueness, which can

be disruptive especially in critical areas of deployment such as self-driving cars and biometric identification.
  - Resource Constraints: Consequently, GAN training requires large computational and data power that makes them less accessible and less scalable, particularly in today's constrained environments.
  - Explicable GANs: Creating models to improve the level of transparence and interpretation of GAN based results.
  - Ethical Guidelines: Setting up international benchmarks to ensure that GAN use is compliant with privacy and security laws.
- Efficiency Improvements: Developing new methods for constructing GANs of low complexity and simplified forms that allow their deployment.

Call to action for interdisciplinary research

To realize the full potential of GANs in information security while mitigating associated risks, this review underscores the need for collaborative, interdisciplinary efforts:

- Bridging AI and Security: Strengthen the synergy of AI-related research with the cybersecurity field to architect highly flexible and real time threat prevention systems.
- Policy and Ethical Development: Coordinate with technical and policy stakeholders to develop appropriate innovative control systems to encourage or require proper regulatory measures of GAN to address such duality.
- Exploring Emerging Applications: In order to address new and developing cybersecurity threats, it is necessary to investigate the potential use of GANs within IoT protection, blockchain, and automated incident response systems.
- Further academic work should focus on stabilizing the training of GANs, enhancing computational cost effectiveness, and improving model interpretation. Such future developments will ensure that GANs are associated with a revolutionary leap in the formation of safe and ethically unambiguous cybersecurity systems.

Authors' contributions

Zaid Arafat had the idea and planned the review, did the systematic search on literature, condensed the findings on both the GAN architectures and cybersecurity implementations, and wrote the major part of the manuscript.

Olga V. Yudina was involved in the development of the methodological framework, critically reviewed and reconstructed the manuscript because of significant intellectual content, participated in the development of the aim and the research methodology, and provided senior supervision in the course of the study.

Zainab A. Abdulazeez helped in data curation, tabulated the important studies and comparisons in performance, and participated in the writing and editing of the final paper.

Each of the authors read and gave their approval to the final version of the manuscript.

REFERENCES

1. Goodfellow I., Pouget-Abadie J., Mirza M., et al. Generative adversarial networks. *Commun. ACM.* 2020;63(11):139–144. https://doi.org/10.1145/3422622
2. Arifin M.M., Ahmed M.S., Ghosh T.K., Udoy I.A., Zhuang J., Yeh J. A Survey on the Application of Generative Adversarial Networks in Cybersecurity: Prospective. Direction and Open Research Scopes. 2024. *ArXiv Prepr.* arXiv:2407.08839. https://doi.org/10.48550/arXiv.2407.08839
3. Sabuhi M., Zhou M., Bezemer C.-P., Musilek P. Applications of Generative Adversarial Networks in Anomaly Detection: A Systematic Literature Review. *IEEE Access*. 2021;9:161003–161029. https://doi.org/10.1109/ACCESS.2021.3131949
4. Aggarwal A., Mittal V., Battineni G. Generative adversarial network: An overview of theory and applications. *Int. J. Inf. Manag. Data Insights*. 2021;1(1):100004. https://doi.org/10.1016/j.jjimei.2020.100004
5. Cao Y.-J., Jia L.-L., Chen Y.-X., et al. Recent Advances of Generative Adversarial Networks in Computer Vision. *IEEE Access.* 2019;7:14985–15006. https://doi.org/10.1109/ACCESS.2018.2886814
6. Radford A., Metz L., Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. 2016.*ArXiv Prepr.* arXiv:1511.06434. https://doi.org/10.48550/arXiv.1511.06434
7. Arjovsky M, Chintala S., Bottou L. Wasserstein generative adversarial networks. In: *Proceedings of the International Conference on Machine Learning* (*ICML*). PMLR. 2017. P. 214–223. Available from URL: https://proceedings.mlr.press/v70/arjovsky17a/arjovsky17a.pdf
8. Zhu J.-Y., Park T., Isola P., Efros A.A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. 2020. *ArXiv Prepr.* arXiv:1703.10593. https://doi.org/10.48550/arXiv.1703.10593
9. Mirza M., Osindero S. Conditional Generative Adversarial Nets. 2014. *ArXiv Prepr.* arXiv:1411.1784. https://doi.org/10.48550/arXiv.1411.1784

10. Karras T., Laine S., Aila T. A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). 2019. P. 4401–4410. https://doi.org/10.1109/CVPR.2019.00453
11. Sedjelmaci H. Attacks detection and decision framework based on generative adversarial network approach: Case of vehicular edge computing network. *Trans. Emerg. Telecommun. Technol.* 2022;33(10):e4073. https://doi.org/10.1002/ett.4073
12. Kumaran U., Thangam S., Prabhakar T.N., Selvaganesan J., Vishwas H.N. Adversarial Defense: A GAN-IF Based Cybersecurity Model for Intrusion Detection in Software Piracy. *J. Wirel. Mob. Netw. Ubiquitous Comput. Dependable Appl.* 2023;14(4):96–114. http://doi.org/10.58346/JOWUA.2023.I4.008
13. Haloui I., Gupta J.S., Feuillard V. Anomaly detection with Wasserstein GAN. 2018. *ArXiv Prepr.* arXiv:1812.02463. https://doi.org/10.48550/arXiv.1812.02463
14. Kimura D., Chaudhury S., Narita M., Munawar A., Tachibana R. Adversarial Discriminative Attention for Robust Anomaly Detection. In: *2020 IEEE Winter Conference on Applications of Computer Vision* (*WACV*). IEEE; 2020. P. 2161–2170. https://doi.org/10.1109/WACV45572.2020.9093428
15. Dunmore A., Jang-Jaccard J., Sabrina F., Kwak J. A Comprehensive Survey of Generative Adversarial Networks (GANs) in Cybersecurity Intrusion Detection. *IEEE Access.* 2023;11:76071–76094. https://doi.org/10.1109/ACCESS.2023.3296707
16. Kos J., Fischer I., Song D. Adversarial examples for generative models. 2017. *ArXiv Prepr.* arXiv:1702.06832. https://doi.org/10.48550/arXiv.1702.06832
17. Chhetri S.R., Lopez A.B., Wan J., Al Faruque M.A. GAN-Sec: Generative Adversarial Network Modeling for the Security Analysis of Cyber-Physical Production Systems. In: *2019 Design. Automation & Test in Europe Conference & Exhibition* (*DATE*). IEEE; 2019. P. 770–775. https://doi.org/10.23919/DATE.2019.8715283
18. Mao X., Li Q., Xie H., et al. Least squares generative adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision* (*ICCV*). 2017. P. 2794–2802. https://doi.org/10.1109/ICCV.2017.304
19. Nataraj L., Karthikeyan S., Jacob G., Manjunath B.S. Malware images: visualization and automatic classification. In: *Proceedings of the 8th International Symposium on Visualization for Cyber Security.* 2011. P. 1–7. https://doi.org/10.1145/2016904.2016908
20. Chen X., Duan Y., Houthooft R., et al. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In: *Advances in Neural Information Processing Systems* (*NeurIPS*). 2016;29:2172–2180.
21. Alo S.O., Jamil A.S., Hussein M.J., Al-Dulaimi M.K.H., Taha S.W., Khlaponina A. Automated Detection of Cybersecurity Threats Using Generative Adversarial Networks (GANs). In: *2024 36th Conference of Open Innovations Association* (*FRUCT*). IEEE. 2024. P. 566–577. https://doi.org/10.23919/FRUCT64283.2024.10749874
22. Zhang J., Li C. Adversarial examples: Opportunities and challenges. *IEEE Trans. Neural Netw. Learn. Syst.* 2019;31(7):2578–2593. https://doi.org/10.1109/TNNLS.2019.2933524
23. Zhang S., Xie X., Xu Y. A Brute-Force Black-Box Method to Attack Machine Learning-Based Systems in Cybersecurity. *IEEE Access*. 2020;8:128250–128263. https://doi.org/10.1109/ACCESS.2020.3008433
24. Papernot N., McDaniel P., Goodfellow I., Jha S., Celik Z.B., Swami A. Practical Black-Box Attacks against Machine Learning. In: *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM. 2017. P. 506–519. https://doi.org/10.1145/3052973.3053009
25. Kurakin A., Goodfellow I.J., Bengio S. Adversarial examples in the physical world. In book: *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC. 2018. P. 99–112. https://doi.org/10.1201/9781351251389, Available from URL: https://www.taylorfrancis.com/chapters/edit/10.1201/9781351251389-8/adversarial-examples-physical-world-alexey-kurakin-ian-goodfellow-samy-bengio
26. Taheri S., Khormali A., Salem M., Yuan J.-S. Developing a robust defensive system against adversarial examples using generative adversarial networks. *Big Data Cogn. Comput.* 2020;4(2):11. https://doi.org/10.3390/bdcc4020011
27. Carlini N., Wagner D. Towards evaluating the robustness of neural networks. In: *2017 IEEE Symposium on Security and Privacy* (*SP*). IEEE. 2017. P. 39–57. https://doi.org/10.1109/SP.2017.49
28. Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A. Towards Deep Learning Models Resistant to Adversarial Attacks. 2019. *ArXiv Prepr.* arXiv:1706.06083. https://doi.org/10.48550/arXiv.1706.06083
29. Sharif M., Bhagavatula S., Bauer L., Reiter M.K. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2016. P. 1528–1540. https://doi.org/10.1145/2976749.2978392
30. Akhtar N., Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*. 2018;6:14410–14430. https://doi.org/10.1109/ACCESS.2018.2807385
31. Dong Y., Pang T., Su H., Zhu J. Evading defenses to transferable adversarial examples by translation-invariant attacks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019. P. 4312–4321. Available from URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Dong_Evading_Defenses_to_Transferable_Adversarial_Examples_by_Translation-Invariant_Attacks_CVPR_2019_paper.html
32. Shafahi A., Najibi M., Ghiasi A., et al. Adversarial training for free! *Adv. Neural Inf. Process. Syst.* 2019:32. Available from URL: https://proceedings.neurips.cc/paper/by-source-2019-1853
33. Xiao C., Li B., Zhu J., He W., Liu M., Song D. Generating Adversarial Examples with Adversarial Networks. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence.* International Joint Conferences on Artificial Intelligence Organization. 2018. P. 3905–3911. https://doi.org/10.24963/ijcai.2018/543

34. Hou T., Wang T., Lu Z., Liu Y., Sagduyu Y. IoTGAN: GAN powered camouflage against machine learning based IoT device identification. In: *2021 IEEE International Symposium on Dynamic Spectrum Access Networks* (*DySPAN*). IEEE. 2021. P. 280–287. https://doi.org/10.1109/DySPAN53946.2021.9677264
35. Kurakin A., Goodfellow I., Bengio S. Adversarial examples in the physical world. 2017. *ArXiv Prepr.* arXiv:1607.02533. https://doi.org/10.48550/arXiv.1607.02533.
36. Goodfellow I., Pouget-Abadie J., Mirza M., et al. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 2014;27. Available from URL: https://proceedings.neurips.cc/paper/5423-generative-adversarial-nets
37. Bengio Y. Learning Deep Architectures for AI. *Found. Trends® Mach. Learn.* 2009;2(1):1–127. https://doi.org/10.1561/2200000006
38. Isola P., Zhu J.-Y., Zhou T., Efros A.A. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 2017. P. 1125–1134. Available from URL: http://openaccess.thecvf.com/content_cvpr_2017/html/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.html
39. Salimans T., Goodfellow I., Zaremba W., et al. Improved techniques for training GANs. In: *Adv. Neural Inf. Process. Syst.* (*NeurIPS*). 2016;29. Available from URL: https://proceedings.neurips.cc/paper_files/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html
40. Arjovsky M., Bottou L. Towards Principled Methods for Training Generative Adversarial Networks. 2017. *ArXiv Prepr.* arXiv:1701.04862. https://doi.org/10.48550/arXiv.1701.04862
41. Ho J., Ermon S. Generative adversarial imitation learning. *Adv. Neural Inf. Process. Syst.* (*NeurIPS*). 2016;29. Available from URL: https://papers.nips.cc/paper_files/paper/2016/hash/cc7e2b878868cbae992d1fb743995d8f-Abstract.html
42. Mittal S., Joshi A., Finin T. Cyber-All-Intel: An AI for Security related Threat Intelligence. 2019. *ArXiv Prepr.* arXiv:1905.02895. https://doi.org/10.48550/arXiv.1905.02895
43. Yinka-Banjo C., Ugot O.-A. A review of generative adversarial networks and its application in cybersecurity. *Artif. Intell. Rev.* 2020;53(3):1721–1736. https://doi.org/10.1007/s10462-019-09717-4
44. Yan Q., Wang M., Huang W., Luo X., Yu F.R. Automatically synthesizing DoS attack traces using generative adversarial networks. *Int. J. Mach. Learn. Cybern.* 2019;10(12):3387–3396. https://doi.org/10.1007/s13042-019-00925-6
45. Goodfellow I.J., Pouget-Abadie M., Mirza M., et al. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* (*NeurIPS*). 2014;27. Available from URL: https://papers.nips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html
46. Choi Y., Choi M., Kim M., Ha J.-W., Kim S., Choo J. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE. 2018. P. 8789–8797. https://doi.org/10.1109/CVPR.2018.00916
47. Karras T., Aila T., Laine S., Lehtinen J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In: *International Conference on Learning Representations*. 2018. Available from URL: https://research.aalto.fi/en/publications/progressive-growing-of-gans-for-improved-quality-stability-and-va
48. Brock A., Donahue J., Simonyan K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In: *International Conference on Learning Representations*. 2018. https://doi.org/10.48550/arXiv.1809.11096
49. Wang T.-C., Liu M.-Y., Zhu J.-Y., Tao A., Kautz J., Catanzaro B. High-resolution image synthesis and semantic manipulation with conditional gans. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 2018. P. 8798–8807. Available from URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_High-Resolution_Image_Synthesis_CVPR_2018_paper.html
50. Li C., Wand M. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. In: Leibe B., Matas J., Sebe N., Welling M. (Eds.). *Computer Vision – ECCV 2016*. Series: Lecture Notes in Computer Science. Cham: Springer; 2016. V. 9907. P. 702–716. https://doi.org/10.1007/978-3-319-46487-9_43
51. Mirsky Y., Lee W. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* Jan. 2022;54(1):1–41. https://doi.org/10.1145/3425780
52. Odena A., Olah C., Shlens J. Conditional image synthesis with auxiliary classifier GANs. In: *Proceedings of the 34th International Conference on Machine Learning* (*ICML*). PMLR. 2017. P. 2642–2651. Available from URL: https://proceedings.mlr.press/v70/odena17a.html
53. Wang Z., She Q., Ward T.E. Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy. *ACM Comput. Surv.* 2022;54(2):1–38. https://doi.org/10.1145/3439723
54. Creswell A., White T., Dumoulin V., Arulkumaran K., Sengupta B., Bharath A.A. Generative Adversarial Networks: An Overview. *IEEE Signal Process. Mag.* 2018;35(1):53–65. https://doi.org/10.1109/MSP.2017.2765202
55. Zhang H., Goodfellow I., Metaxas L., et al. Self-attention generative adversarial networks. In: *Proceedings of the 36th International Conference on Machine Learning* (*ICML*). PMLR. 2019. P. 7354–7363. Available from URL: https://proceedings.mlr.press/v97/zhang19d.html
56. Lucic M., Kurach K., Michalski M., Gelly S., Bousquet O. Are gans created equal? A large-scale study. *Adv. Neural Inf. Process. Syst.* (*NeurIPS*) 2018;31. Available from URL: https://proceedings.neurips.cc/paper/2018/hash/e46de7e1bcaaced9a54f1e9d0d2f800d-Abstract.html
57. Sun H., Zhu T., Zhang Z., Xiong D.J.P., Zhou W. Adversarial Attacks Against Deep Generative Models on Data: A Survey. *IEEE Trans. Knowl. Data Eng.* 2023;35(4):3367–3388. https://doi.org/10.1109/TKDE.2021.3130903
58. Miyato T., Kataoka T., Koyama M., Yoshida Y. Spectral Normalization for Generative Adversarial Networks. 2018. *ArXiv Prepr.* arXiv:1802.05957. https://doi.org/10.48550/arXiv.1802.05957

59. Che T., Li Y., Jacob A.P., Bengio Y., Li W. Mode Regularized Generative Adversarial Networks. 2017. *ArXiv Prepr.* arXiv:1612.02136. https://doi.org/10.48550/arXiv.1612.02136
60. Bao J., Chen D., Wen F., Li H., Hua G. CVAE-GAN: Fine-Grained Image Generation Through Asymmetric Training. *Presented at the Proceedings of the IEEE International Conference on Computer Vision* (*ICCV*). 2017. P. 2745–2754. Available from URL: https://openaccess.thecvf.com/content_iccv_2017/html/Bao_CVAE-GAN_Fine-Grained_Image_ICCV_2017_paper.html
61. Westerlund M. The emergence of deepfake technology: A review. *Technol. Innov. Manag. Rev.* 2019;9(11):39–52.
62. Tolosana R., Vera-Rodriguez R., Fierrez J., Morales A., Ortega-Garcia J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion*. 2020;64:131–148. https://doi.org/10.1016/j.inffus.2020.06.014
63. Zhao J. Energy-based Generative Adversarial Network. 2016. *ArXiv Prepr.* arXiv:1609.03126. https://doi.org/10.48550/arXiv.1609.03126
64. Reed S., Akata Z., Yan X., Logeswaran L., Schiele B., Lee H. Generative adversarial text to image synthesis. In: *Proceedings of the 33th International Conference on Machine Learning*. PMLR. 2016. P. 1060–1069. Available from URL: http://proceedings.mlr.press/v48/reed16.html
65. Lloyd S., Weedbrook C. Quantum Generative Adversarial Learning. *Phys. Rev. Lett.* 2018;121(4):040502. https://doi.org/10.1103/PhysRevLett.121.040502
66. Gulrajani I., Ahmed F., Arjovsky M., Dumoulin V., Courville A.C. Improved training of wasserstein gans. *Adv. Neural Inf. Process. Syst.* 2017;30. Available from URL: https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html
67. Park T., Liu M.-Y., Wang T.-C., Zhu J.-Y. Semantic image synthesis with spatially-adaptive normalization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019. P. 2337–2346. Available from URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Park_Semantic_Image_Synthesis_With_Spatially-Adaptive_Normalization_CVPR_2019_paper.html
68. Hoang Q., Nguyen T.D., Le T., Phung D. *MGAN: Training Generative Adversarial Nets with Multiple Generators*. 2018.
69. Hitaj B., Gasti P., Ateniese G., Perez-Cruz F. PassGAN: A Deep Learning Approach for Password Guessing. 2019. *ArXiv Prepr.* arXiv:1709.00440. https://doi.org/10.48550/arXiv.1709.00440
70. Sharma Y., Ding G.W., Brubaker M. On the Effectiveness of Low Frequency Perturbations. 2019. *ArXiv Prepr.* arXiv:1903.00073. https://doi.org/10.48550/arXiv.1903.00073
71. Zhang C., Yu S., Tian Z., Yu J.J.Q. Generative Adversarial Networks: A Survey on Attack and Defense Perspective. *ACM Comput. Surv.* 2024;56(4):1–35. https://doi.org/10.1145/3615336
72. Zhang J., Zhao L., Yu K., Min G., Al-Dubai A.Y., Zomaya A.Y. A Novel Federated Learning Scheme for Generative Adversarial Networks. *IEEE Trans. Mob. Comput.* 2024;23(5):3633–3649. https://doi.org/10.1109/TMC.2023.3278668
73. Kaviani S., Han K.J., Sohn I. Adversarial attacks and defenses on AI in medical imaging informatics: A survey. *Expert Syst. Appl.* 2022;198:116815. https://doi.org/10.1016/j.eswa.2022.116815
74. Ribeiro M.T., Singh S., Guestrin C. Why Should I Trust You? Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA. 2016. P. 1135–1144. https://doi.org/10.1145/2939672.2939778
75. Zhang Q., Wu Y.N., Zhu S.-C. Interpretable convolutional neural networks. In: *Proceedings of the IEEE/CVPR Conference on Computer Vision and Pattern Recognition*. 2018. P. 8827–8836. https://doi.org/10.1109/CVPR.2018.00920
76. Borji A. Pros and Cons of GAN Evaluation Measures. 2018. *ArXiv Prepr.* arXiv:1802.03446. https://doi.org/10.48550/arXiv.1802.03446
77. Yang Y., Li Y., Zhang W., Qin F., Zhu P., Wang C.-X. Generative-Adversarial-Network-Based Wireless Channel Modeling: Challenges and Opportunities. *IEEE Commun. Mag.* 2019;57(3):22–27. https://doi.org/10.1109/MCOM.2019.1800635
78. Li T., Zhang S., Xia J. Quantum generative adversarial network: A survey. *Comput. Mater. Contin.* 2020;64(1):401–438. https://doi.org/10.32604/cmc.2020.010551
79. Zhao S., Liu Z., Lin J., Zhu J.-Y., Han S. Differentiable augmentation for data-efficient GAN training. *Adv. Neural Inf. Process. Syst.* 2020;33:7559–7570. Available from URL: https://proceedings.neurips.cc/paper/2020/hash/55479c55ebd1efd3ff125f1337100388-Abstract.html
80. Mittelstadt B., Russell C., Wachter S. Explaining Explanations in AI. In: *Proceedings of the Conference on Fairness. Accountability. and Transparency*. Atlanta, GA, USA. 2019. P. 279–288. https://doi.org/10.1145/3287560.3287574

About the Authors

Zaid Arafat, Assistant Lecturer, Department of Cybersecurity, University of Kerbala (Karbala, 56001 Iraq). E-mail: zaid.q@uokerbala.edu.iq. Scopus Author ID 57963547500, https://orcid.org/0009-0001-0886-5370

Olga V. Yudina, Cand. Sci. (Eng.), Associate Professor, Department of Mathematics and Computer Software, Cherepovets State University (5, Lunacharskogo pr., Cherepovets, 162600 Russia). E-mail: oviudina@chsu.ru. RSCI SPIN-code 7741-5343, https://orcid.org/0009-0005-6367-1076

Zainab A. Abdulazeez, Assistant Lecturer, College of Education for Human Sciences, University of Kerbala (Karbala, 56001 Iraq). E-mail: zainab.abdulhameed@uokerbala.edu.iq. Scopus Author ID 57220186609, https://orcid.org/0009-0004-9801-4888

Об авторах

Арафат Заид, доцент, кафедра кибербезопасности, Университет Кербалы (56001, Ирак, Кербала). E-mail: zaid.q@uokerbala.edu.iq. Scopus Author ID 57963547500, https://orcid.org/0009-0001-0886-5370

Юдина Ольга Вадимовна, к.т.н., доцент, доцент кафедры математического и программного обеспечения ЭВМ, ФГБОУ ВО «Череповецкий государственный университет» (162600, Россия, Череповец, пр-т Луначарского, д. 5). E-mail: oviudina@chsu.ru. SPIN-код РИНЦ 7741-5343, https://orcid.org/0009-0005-6367-1076

Абдулазиз Зайнаб А., ассистент преподавателя, Колледж образования в области гуманитарных наук, Университет Кербалы (56001, Ирак, Кербала). E-mail: zainab.abdulhameed@uokerbala.edu.iq. Scopus Author ID 57220186609, https://orcid.org/0009-0004-9801-4888

The text was submitted by the authors in English
Edited for English language and spelling by Thomas A. Beavitt