INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY FUTURISTIC DEVELOPMENT

Review of Hybrid Generative–Discriminative Models in Cybersecurity

Hadeel Mohsen Ibrahim
Ministry of Education Directorate of Education Rusafa-2, Baghdad, Iraq

* Corresponding Author: **Hadeel Mohsen Ibrahim**

Article Info

P-ISSN: 3051-3618
E-ISSN: 3051-3626
Volume: 06
Issue: 02
July - December 2025
Received: 05-09-2025
Accepted: 06-10-2025
Published: 15-11-2025
Page No: 86-106

Abstract
Hybrid generative–discriminative models have become a strong model in cybersecurity because they combine strengths that help solve problems that have been around for a long time, such as detection accuracy, data imbalance, adversarial robustness, and zero-day generalization. This review combines new work that combines generative models like variational autoencoders, generative adversarial networks, diffusion models, and probabilistic graphical models with discriminative models like convolutional networks, recurrent and transformer-based models, graph neural networks, and traditional machine-learning classifiers. The research offers a systematic classification of hybrid methodologies, emphasizing joint, sequential, parallel, and adversarially focused frameworks. It also looks at how well they work in important areas of cybersecurity, such as finding intrusions, analyzing malware, spotting fraud and anomalies, classifying phishing and spam, and defending against attacks. Key performance insights show that hybrid models are always better than standalone classifiers at recognizing minority classes, representing quality, and being able to handle new threats. Even though they hold a lot of promise, there are still big problems with scalability, explainability, data privacy, and defense against adaptive attackers. This review gives a complete picture of the current limitations and future research directions for the next generation of intelligent, resilient, and trustworthy cybersecurity systems. These include multimodal foundation models, continual learning, privacy-preserving hybrid pipelines, and standardized benchmarks.

Keywords: Hybrid Models, Generative Learning, Cybersecurity, Intrusion Detection, Adversarial Robustness, Deep Learning

1. Introduction

Cybersecurity has become a very important part of our social and technical infrastructure because attacks are more and more often targeting cloud platforms, industrial control systems, IoT ecosystems, and important services. Conventional rule-based and exclusively discriminative machine learning (ML) approaches frequently encounter challenges related to concept drift, data imbalance, zero-day attacks, and adversarial manipulation [1, 2]. Recent advancements in generative models—such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and associated architectures—have created novel opportunities for data-centric security solutions, including synthetic attack generation, traffic simulation, and representation learning. Systematic surveys have recorded the utility of generative models in intrusion detection, penetration testing, and dataset augmentation; however, the majority of these studies examine generative and discriminative models in isolation rather than as integrated hybrids. [3, 4]

In parallel, the machine learning community has a long tradition of hybrid generative–discriminative models, where generative components capture data structure while discriminative components optimize predictive performance. Early work by Raina *et al*. proposed hybrid models that interpolate between naive Bayes and logistic regression, demonstrating improved performance when labeled data are scarce [5, 1] later applied hybrid generative–discriminative hidden Markov models for activity recognition in sensor-rich environments, showing that such hybrids can better exploit temporal and structural regularities than either family alone [6].

Despite clear conceptual alignment between these hybrid principles and the needs of cybersecurity—where unlabeled data are abundant but labels are costly and attack distributions evolve rapidly—there is no focused review that systematically analyzes hybrid generative–discriminative models in cybersecurity as a distinct methodological class. Existing reviews tend to focus either on GANs alone, on generative models broadly, or on hybrid *discriminative* ensembles with little emphasis on explicit generative components. [7] This paper addresses that gap.

1.1. Background and Motivation

Generative models learn to approximate the joint distribution p(x, y) or the data distribution p(x). This lets them make realistic samples, find hidden structures, and model uncertainty. Discriminative models, on the other hand, try to find the conditional distribution p(y|x) or learn direct decision boundaries. When there is enough labeled data, this usually leads to better predictive accuracy. Theoretical and empirical research indicates that generative models can surpass discriminative models in low-label contexts, whereas discriminative models excel in scenarios with abundant labels. [8] This complementarity drives hybrid strategies that aim to integrate the advantages of both families. In cybersecurity, real-world problems fit well with the strengths of generative modeling. IDS and malware-detection datasets frequently exhibit significant imbalance, characterized by underrepresentation of minority attack classes in comparison to benign traffic [9]. Recent research indicates that GAN-based augmentation and VAE-based synthesis can efficiently rebalance datasets and enhance classifier performance in network intrusion detection and industrial Internet contexts. [10, 11] Halvorsen *et al*. also say that generative machine learning models (GMLMs) are being used more and more not only to add more data but also as stand-alone detectors and tools for penetration testing in intrusion detection systems (IDSs). [12]

Simultaneously, advanced discriminative architectures—such as CNNs, RNNs/LSTMs, transformers, graph neural networks, and gradient-boosted trees—have gained extensive utilization in intrusion detection, malware classification, fraud detection, and anomaly detection [13]. A new hybrid ML–DL IDS framework that uses XGBoost, CNN, and LSTM showed strong results on several benchmark datasets (CIC-IDS2017, UNSW-NB15, NSL-KDD, WSN-DS). [14] Complementary systematic reviews underscore the challenges in attaining robust, real-time, and adversarially resilient IDS performance, pinpointing data scarcity, class imbalance, and non-stationary attack distributions as principal impediments. [15] These trends drive the development of hybrid generative-discriminative architectures, where generative models yield structured representations, synthetic samples, or uncertainty estimates, and discriminative models enhance classification performance. Because these kinds of approaches are growing so quickly, there is a clear need for a focused review that brings together the scattered literature and looks at how hybrid designs help with cybersecurity tasks in a real way.

1.2. Role of Hybrid Generative–Discriminative Models in Cybersecurity

Hybrid generative–discriminative models occupy a methodological middle ground: generative components (e.g., VAEs, GANs, graph-VAEs) model complex data distributions or latent spaces [16], while discriminative components (e.g., random forests, deep neural networks) exploit those representations for classification, regression, or ranking. In early hybrid models, generative priors were combined with discriminative training for parts of the parameter space, leading to better generalization under limited labeled data. [17, 18, 19]

In cybersecurity, this idea appears in several concrete patterns:

1. **Generative representation learning + discriminative classifier.** Ajayi *et al*. propose learning low-dimensional latent representations of malware with a VAE and then training conventional classifiers (decision trees, logistic regression, LightGBM, random forests) on the latent features, demonstrating comparable or improved detection performance with reduced computational cost. [20] Graph-based VAEs have also been used to compress API-call graphs for malware detection, enhancing performance while reducing feature dimensionality. [21]
2. **Generative data augmentation + discriminative IDS.** GAN and Wasserstein-GAN variants are increasingly used to generate synthetic attack traffic and rebalance minority classes in industrial Internet and network IDS datasets. Feng *et al*. introduce a SA-WGAN model that enhances industrial Internet intrusion detection by generating high-quality, diverse samples for underrepresented classes, which are then consumed by discriminative classifiers. [22] Similar GAN-based augmentation strategies are reported to improve IoT IDS performance and mitigate data imbalance in benchmark datasets such as NSL-KDD. [23]
3. **Hybrid sampling or representation pipelines.** Guo and Xie propose a network intrusion detection architecture that combines hybrid sampling methods with deep learning classifiers, addressing class imbalance and improving detection rates. [24] Dong *et al*. develop a VAE-based abnormal traffic detection algorithm, which can be integrated with discriminative decision layers for anomaly scoring in network traffic. [25]
4. **System-level integration of generative IDS components.** Halvorsen *et al*. conduct a systematic mapping study on generative ML for intrusion detection, identifying three main application areas: (i) using generative models as IDSs, (ii) supplementing IDS datasets, and (iii) supporting penetration testing and attack simulation. [26] These application areas provide natural interfaces between generative components and discriminative IDS pipelines.

Together, these examples show that hybrid generative–discriminative models are not a single architecture but a design paradigm that can be instantiated in multiple ways: data-level hybrids (synthetic data + classifier), representation-level hybrids (latent features + classifier), and training-level hybrids (joint or alternating optimization of generative and discriminative objectives). A structured review is needed to organize these patterns and evaluate their benefits and limitations across different cybersecurity tasks [27, 28].

1.3. Problem Statement and Research Gap

While the literature on generative models in cybersecurity has expanded rapidly, existing surveys remain fragmented along either model-type or task-type boundaries. Al-Ajlan

and Ykhlef review GAN-based methods specifically for intrusion detection systems, focusing on architectures, datasets, and evaluation metrics but largely restricting attention to GAN–IDS combinations. [3] Arifin *et al*. offer a broader survey of GAN applications in cybersecurity—including malware detection, anomaly detection, and intrusion detection—but still treat GANs as the central paradigm and do not systematically analyze integration with discriminative classifiers. [4]

Halvorsen *et al*. provide a comprehensive mapping of generative machine learning models in intrusion detection, highlighting how GMLMs can supplement datasets, generate attack flows, and sometimes serve as primary detectors. [5] However, their focus is on generative models themselves; discriminative classifiers are treated primarily as baselines or downstream components, and hybrid design principles are not the central topic. Similarly, recent systematic reviews on IDS technologies and deep learning approaches emphasize challenges like data imbalance, real-time constraints, and adversarial robustness but do not frame hybrid generative–discriminative modeling as an organizing concept. [29, 30]

On the other hand, classical ML works on hybrid generative–discriminative models concentrate on generic pattern recognition problems (text classification, activity recognition, facial expression recognition) and predate the current wave of deep generative models and cybersecurity applications. [1, 2] As a result, there is currently no integrative review that:

- Treats hybrid generative–discriminative architectures as the *primary unit of analysis* in cybersecurity;
- Compares different hybrid patterns (data-level, representation-level, and training-level hybrids);
- Evaluates how these hybrids address concrete cybersecurity challenges such as zero-day detection, class imbalance, and adversarial robustness.

This paper addresses this gap by providing a systematic and conceptually grounded review of hybrid generative–discriminative models in cybersecurity.

1.4. Objectives of the Review

The primary aim of this review is to systematically consolidate the current advancements in hybrid generative–discriminative models within cybersecurity and to establish a cohesive conceptual framework for this developing field. The review has the following specific goals:

1. **Conceptual clarification** – To delineate the characteristics of a hybrid generative–discriminative model within the cybersecurity domain, differentiating it from exclusively generative or exclusively discriminative approaches, as well as from general "hybrid" ML–DL ensembles. [32]
2. **Taxonomic organization** – To create a taxonomy of hybrid architectures (for example, generative representation + classifier, generative augmentation + classifier, joint hybrid training) and fit current cybersecurity studies into this taxonomy. [33]
3. **Critical performance assessment** – To combine real-world data on detection accuracy, false-positive rates, robustness, and computational overhead with non-hybrid baselines, with a focus on data imbalance and how well the model works with new attacks. [34]
4. **Identification of open challenges and research directions** – To point out methodological and practical gaps, like not having standardized benchmarks, not being able to interpret results well, and not being able to evaluate results well in adversarial settings, and to suggest specific ways for future research to go. [5]

1.5. Scope, Limitations, and Contribution of This Study

Scope.

This review focuses on peer-reviewed studies in which both a generative and a discriminative component play a substantive role in the cybersecurity pipeline. We include works where generative models are used for latent representation learning or data augmentation and where their outputs directly feed discriminative classifiers for tasks such as intrusion detection, malware classification, fraud detection, or more specialized cybersecurity problems (e.g., cyber-physical systems, industrial Internet). [35, 36] Purely generative adversarial attack methods without an explicit defensive discriminative component are considered only when they form part of a hybrid training or evaluation scheme.

Exclusions. We do *not* attempt to provide an exhaustive survey of:

- All GAN-based security applications (these are already covered by dedicated surveys [36]);
- All deep learning–based IDS or malware detection frameworks that do not involve an explicit generative model [37];
- Classical statistical anomaly detection methods without generative modeling components.

Methodological limitations.

Although we leverage recent systematic reviews and mapping studies as starting points, [38] the hybrid nature of many works is often implicit rather than explicitly labeled. Consequently, some studies that could be interpreted as hybrid may not be captured if their generative or discriminative components are not clearly described. Moreover, the rapid pace of publication, particularly in conferences and preprints, means that any snapshot will be incomplete.

Contributions

Within these boundaries, this review makes three primary contributions:

1. It articulates a clear definition and taxonomy of hybrid generative–discriminative models tailored to cybersecurity, bridging classical hybrid modeling concepts [11, 7] with modern deep generative architectures.
2. It maps and critically synthesizes recent cybersecurity studies that exemplify this hybrid paradigm, encompassing intrusion detection, malware detection, and associated domains, while analyzing their empirical strengths and weaknesses in comparison to non-hybrid baselines. [39]
3. It finds open research problems and suggests a structured research agenda to move the field of hybrid generative–discriminative cybersecurity models forward. This includes standardized benchmarks, testing in adversarial and distribution-shift scenarios, interpretable hybrid architectures, and deployment considerations. [40]

These contributions collectively seek to furnish researchers and practitioners with a cohesive framework for the design, assessment, and enhancement of hybrid generative–discriminative models that can adapt to the dynamic requirements of contemporary cybersecurity.

2. Foundations of Generative and Discriminative Modeling

Before looking at how generative and discriminative modeling can be combined in cybersecurity, you need to have a deep understanding of both. Generative models seek to encapsulate the joint distribution p(x, y) or the marginal p(x), while discriminative models concentrate on the conditional distribution p(y|x) or directly acquire a decision function f(x) that correlates inputs to labels. This differentiation has been institutionalized in seminal research juxtaposing logistic regression (discriminative) with naive Bayes (generative) and further elaborated in contemporary theoretical analyses and instructional materials on deep generative models [41, 42]. The decision between generative and discriminative modeling in cybersecurity is not solely technical; it influences how systems manage data scarcity, class imbalance, concept drift, and adaptive adversaries. Recent studies on generative adversarial networks (GANs) and generative AI for intrusion detection indicate that these modeling paradigms are being increasingly integrated in practice, necessitating a solid conceptual framework. [43]

2.1. Definition and Core Principles of Generative Models

Generative models seek to explain how data are produced by learning the joint probability distribution $p(x, y)$or the data distribution $p(x)$, often through latent variables $z$such that $p(x) = \int p(x \mid z)p(z)\, dz$. Classical examples include Gaussian mixture models, hidden Markov models, and naive Bayes. In the deep learning era, deep generative models (DGMs) extend this idea using neural networks as flexible function approximators. Key families include:

- **Explicit density models**, such as normalizing flows and certain autoregressive models, which provide tractable likelihoods.
- **Approximate density models**, notably variational autoencoders (VAEs), which optimize a variational lower bound on the log-likelihood.
- **Implicit density models**, such as GANs, which define a generator implicitly through a sampling process rather than an explicit likelihood. [3, 4, 10]

In a supervised setting, generative classifiers estimate $p(x \mid y)$and $p(y)$, then apply Bayes' rule to obtain $p(y \mid x)$. This enables them to generate synthetic samples conditioned on class labels and to support semi-supervised learning, anomaly detection, and uncertainty-aware inference [44].

Recent theoretical work has revisited the statistical behavior of generative classifiers at scale. Ng and Jordan showed that naive Bayes can reach its asymptotic error with far fewer samples than logistic regression, despite having a higher asymptotic error in the limit of infinite data. [1] generalized these results to multiclass settings and deep representation learning, showing that generative classifiers (e.g., naive Bayes in representation space) can converge faster in few-shot regimes while discriminative models dominate when labeled data are abundant. [5]

In cybersecurity, generative models are used to simulate attack traffic, reconstruct benign behavior, and augment minority attack classes, especially via GAN variants and VAEs. [6–9] This data-centric capability underpins many hybrids generative–discriminative designs.

2.2. Definition and Core Principles of Discriminative Models

Discriminative models aim directly at predicting labels from inputs by modeling the conditional distribution $p(y \mid x)$or learning a decision boundary $f(x)$that separates classes. Logistic regression, support vector machines (SVMs), conditional random fields, and modern deep neural networks (CNNs, RNNs, transformers) are all examples of discriminative models [45].

In the probabilistic view, discriminative models bypass modeling the input distribution $p(x)$and instead maximize the conditional log-likelihood

$$\max_{\theta} \sum_{i=1}^{N} \log p_\theta(y_i \mid x_i), \quad (1)$$

often with regularization to control complexity. This generally yields better asymptotic classification performance than generative models under correct specification because the model capacity is focused on the decision boundary rather than on modeling the entire data-generating process. [22, 15]. Ng and Jordan's comparison between logistic regression and naive Bayes showed two "regimes": with limited training data, naive Bayes (generative) may perform better due to stronger modeling assumptions that reduce variance; with sufficient data, logistic regression (discriminative) overtakes it due to lower asymptotic error. [16] Gwern Subsequent work revisiting this question in modern deep learning contexts has confirmed similar patterns: generative classifiers can converge faster in high-dimensional representation spaces, but discriminative classifiers achieve superior performance given enough labeled samples [5, 11, 12]. Discriminative models are the best at intrusion detection, malware classification, spam filtering, and fraud detection in cybersecurity. They often have the best detection rates on benchmark datasets. Recent surveys of intrusion detection systems (IDSs) stress that ensemble and deep discriminative architectures are the backbone of modern IDSs. However, they also point out problems like overfitting to training distributions, being vulnerable to adversarial examples, and not working well when there is a lot of class imbalance. [8, 9, 13]

2.3. Theoretical Differences and Complementarity

The generative–discriminative distinction can be framed along several theoretical axes:

1. **Modeling target**
   - **Generative:** learns $p(x, y)$or $p(x \mid y)p(y)$, supporting sampling and likelihood evaluation.
   - **Discriminative:** learns $p(y \mid x)$or a decision function $f(x)$, focusing solely on predictive performance. [2, 3, 14]
2. **Sample complexity and asymptotics** Ng and Jordan showed that naive Bayes (generative) can reach its asymptotic error using $O(\log d)$samples in the feature dimension d, while logistic regression (discriminative) requires $O(d)$samples; however, the asymptotic error of logistic regression is lower when data are plentiful. [46] extended this result to multiclass settings and deep representations, again observing a two-regime phenomenon: generative classifiers dominate in low-data regimes, discriminative ones in high-data regimes. [5]

3. **Assumptions and inductive bias.** Generative models typically require stronger assumptions about feature distributions and conditional independence (e.g., naive Bayes), which can be both a strength (when approximately true) and a weakness (when violated). Discriminative models are more flexible but can be data-hungry and less robust when data are scarce or highly imbalanced. [2, 15]
4. **Uncertainty and priors.** Recent work has analyzed generative and discriminative modeling **under the lens of uncertainty**, emphasizing that generative models expose explicit priors and likelihoods, while discriminative models encode priors implicitly. This has implications for epistemic uncertainty, calibration, and robustness under dataset shift. [15]

These differences make the two paradigms complementary rather than mutually exclusive. Hybrid methods—such as discriminative training with generative regularizers, generative pretraining followed by discriminative fine-tuning, or joint models interpolating between the two extremes—can leverage the statistical efficiency of generative models and the asymptotic optimality of discriminative ones. [2, 11, 16] This complementarity is precisely what motivates hybrid architectures in cybersecurity.

2.4. Advantages and Limitations in Cybersecurity Contexts

In cybersecurity, data characteristics strongly influence the suitability of generative and discriminative models: datasets are often imbalanced, noisy, partially labeled, and non-stationary, with evolving attack patterns and active adversaries [46].

Advantages of generative models in cybersecurity include:

- **Data augmentation and class balancing:** GANs and related models can synthesize realistic attack traffic to rebalance heavily skewed datasets, improving downstream discriminative IDS performance. [6, 7, 9, 17]
- **Unsupervised anomaly detection:** VAEs and other DGMs can learn normal behavior and flag deviations as anomalies, which is valuable for zero-day attack detection when labeled attack data are scarce. [4, 10]
- **Threat simulation and testing:** Generative AI can emulate realistic attack sequences or network flows to stress-test IDSs and security policies, an emerging trend in generative AI–based IDS research. [8, 9]

However, generative models also have limitations: their training can be unstable (e.g., GAN mode collapse), generated samples may not capture rare but critical attack behaviors, and complex DGMs can be computationally expensive to deploy in real time. [6, 7]

Advantages of discriminative models in cybersecurity are:

- **High detection accuracy on known patterns**, especially with deep architectures trained on large labeled datasets.
- **Operational efficiency**, as many discriminative models can be optimized for low-latency inference on high-throughput traffic. [13, 18]

Their limitations are equally important: discriminative models can overfit to training distributions, leading to poor generalization under concept drift or novel attack types; they are vulnerable to adversarial examples and may be poorly calibrated (overconfident on out-of-distribution inputs); and their performance often degrades severely under extreme class imbalance. [47]

These complementary strengths and weaknesses motivate hybrid generative–discriminative approaches in cybersecurity: generative components can supply synthetic data, uncertainty estimates, or robust latent representations, while discriminative components provide strong decision boundaries and efficient deployment. Recent surveys on generative AI–based IDSs for IoT and vehicular networks explicitly emphasize such combinations as a promising direction for building adaptive, resilient security solutions [8, 9, 17, 19].

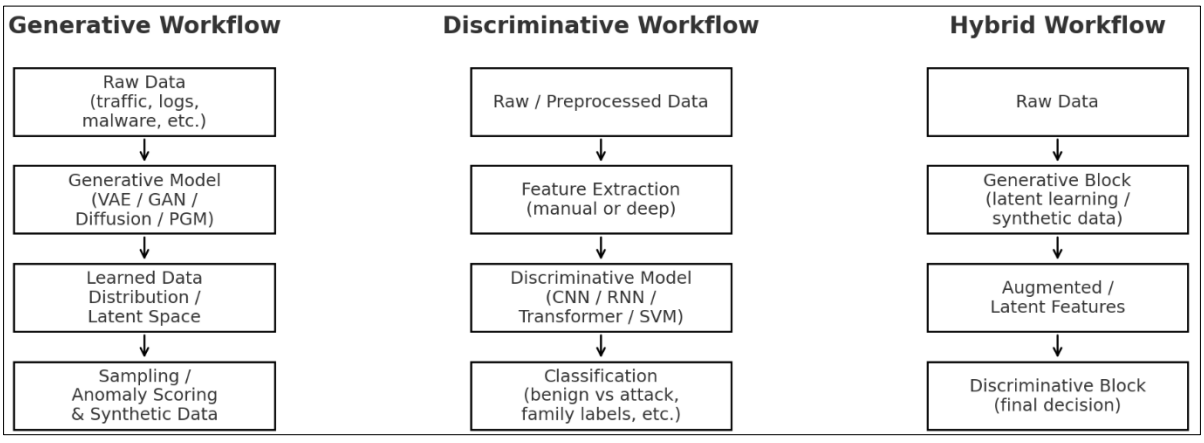

Fig 1: Comparative Workflow of Generative vs Discriminative vs Hybrid Models

3. Hybrid Generative–Discriminative Architectures

Hybrid generative–discriminative architectures are designed to jointly exploit distribution modeling (via generative components) and decision boundary optimization (via discriminative components). Classical work formalized this idea by interpolating between pure generative and pure discriminative training, showing that hybrids can improve robustness in low-data regimes while retaining strong asymptotic performance. [48] In cybersecurity, this design space is instantiated in many forms: VAEs followed by traditional classifiers for malware detection, VAE–GAN data balancers feeding deep IDS models, GAN-based

augmentation coupled with CNN or random forest detectors, and adversarial two-player systems where generators and discriminators are explicitly co-trained. [4–10] To reason about these systematically, it is useful to separate conceptual design, taxonomy, trade-offs, and evaluation criteria.

3.1. Conceptual Design of Hybrid Architectures

Conceptually, a hybrid generative–discriminative system comprises three elements:

1. **Generative component** G_θ: models $p_\theta(x), p_\theta(x \mid y)$, or a latent space $p_\theta(z \mid x)$, often using VAEs, GANs, or hybrids such as VAE–GAN. [49]
2. **Discriminative component** D_ϕ: models $p_\phi(y \mid x)$or a decision function $f_\phi(x)$, typically implemented as classical ML classifiers (SVM, random forest, gradient boosting) or deep networks (CNN, LSTM, transformer). [1–3]
3. **Coupling mechanism**: defines how information flows between $G_\theta$and $D_\phi$and how their objectives are combined (e.g., joint loss, sequential pipeline, ensemble fusion, adversarial game). [2, 3, 6, 11]

Formally, many hybrids optimize a combined objective of the form

$$\mathcal{L}(\theta, \phi) = \alpha\, \mathcal{L}_{\text{gen}}(\theta) + (1 - \alpha)\, \mathcal{L}_{\text{disc}}(\phi, \theta), \quad (2)$$

where $\mathcal{L}_{\text{gen}}$is a generative loss (e.g., negative log-likelihood or VAE evidence lower bound) and $\mathcal{L}_{\text{disc}}$is a discriminative loss (e.g., cross-entropy), with $\alpha \in [0, 1]$controlling the interpolation. This idea underpins principled hybrids of GMMs with discriminative training and general frameworks that reinterpret "discriminative training of generative models" as standard training of an augmented model. [49]

In cybersecurity, design choices depend on the role of the generative block: latent representation learning (e.g., VAE embeddings for malware), synthetic sample generation (e.g., WGAN-augmented IDS), anomaly scoring (e.g., reconstruction error for abnormal traffic), or adversarial training. [4–10, 12] These roles map naturally onto different hybrid patterns described next.

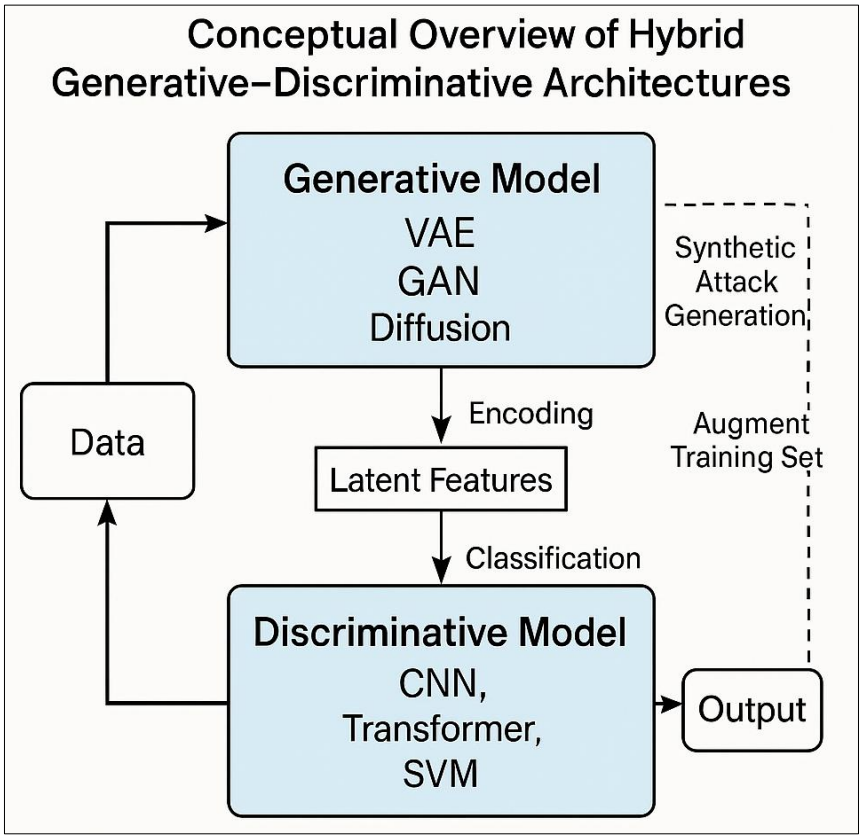

Fig 2: Conceptual Overview of Hybrid Generative–Discriminative Architectures

3.2. Taxonomy of Hybrid Models

3.2.1. Joint Generative–Discriminative Models

Joint hybrids integrate generative and discriminative objectives into a *single* model, typically via a combined loss or augmented probabilistic formulation. Lasserre *et al.* proposed a principled framework in which a generative model is trained with a convex combination of generative and discriminative criteria, yielding models that interpolate between naive Bayes–like and logistic regression–like behavior. [2] Bishop and Lasserre later generalized this viewpoint, showing that "discriminative training of a generative model" can be interpreted as standard maximum-likelihood training of an expanded model. [50]

Recent work extends these ideas to deep and invertible architectures, where flows or deep generative models are trained with both likelihood and classification losses to obtain rich representations and calibrated predictions. [6] Hybrid training of Gaussian mixture models with a discriminative Gaussian-coupled softmax layer is another example, combining mixture modeling of inputs with discriminative class scoring. [11] In cybersecurity, fully joint hybrids are less common but emerging—for example, cross-class VAE–GAN models that simultaneously reconstruct data and optimize class-conditional discrimination for IDS, such as

M2M-VAEGAN-IDS, where VAE–GAN losses are combined with a classifier loss to enhance minority-class intrusion detection. [12]

3.2.2. Sequential Hybrid Models

Sequential hybrids adopt a pipeline structure where G_θ and D_ϕ are trained and applied in sequence. Typical patterns include [51]:

- **Representation first, classification second.** Ajayi *et al*. train a VAE on malware data to obtain low-dimensional latent vectors, then feed these into traditional classifiers (decision tree, naive Bayes, LightGBM, logistic regression, random forest), achieving large dimensionality reduction without sacrificing accuracy. [4]
- **Unsupervised detection followed by supervised refinement.** UL-VAE provides unsupervised zero-day malware detection via a VAE-based anomaly detector, while a supervised classifier handles known families; these two components form a hybrid detection framework for known and novel threats. [5, 13]
- **Data augmentation then discriminative training.** VAE–WGAN-GP and VAE–WACGAN models are trained to generate realistic minority-class attack samples; the augmented datasets are then used to train CNN or other deep IDS classifiers, as demonstrated in industrial Internet intrusion detection and related work. [7, 8]

Sequential hybrids are conceptually simple and easy to retrofit onto existing IDS pipelines: a generative "preprocessor" (for features or data) is followed by a discriminative "backend" (for decisions).

3.2.3. Parallel Hybrid Models

In parallel hybrids, generative and discriminative components operate simultaneously on (possibly the same) inputs, and their outputs are fused at a later stage. Fusion may occur via score averaging, stacking, or meta-learning [52]. bIn computer vision, Alharbi *et al.* propose a hybrid generative–discriminative framework for inpainting forgery detection, where a generative inpainting model and a discriminative classifier are combined to improve robustness against localized manipulations [11]. Similar patterns appear in cybersecurity:

- Hybrid IDS frameworks where an unsupervised autoencoder (or VAE) produces anomaly scores in parallel with a supervised deep classifier, and a meta-layer combines both signals to decide whether a connection is malicious. [9, 10, 14]
- Latent space alignment architectures for IoT botnet detection, where one network learns a generative latent representation and another learns discriminative boundaries; alignment and joint decision rules are then applied to enhance robustness across domains and datasets. [14]

Parallel hybrids are particularly attractive when different components operate on different feature views (e.g., statistical traffic features vs. sequence-based representations) or when diverse uncertainty estimates are needed.

3.2.4. Adversarial Hybrid Models

Adversarial hybrids are built around adversarial training games, typically GAN-style, in which a generator G_θ and a discriminator D_ϕ are trained in opposition. In cybersecurity, these architectures often take one of two forms [53]:

1. **GAN for augmentation + external classifier.** WGAN-DL-IDS uses a WGAN to synthesize minority-class intrusion samples, then trains random forest and deep learning models on the augmented data, improving detection of rare attacks. [9] Similar WGAN-GP and VAE–GAN-based augmentation strategies have been proposed to tackle imbalance in network intrusion datasets, with studies showing that GAN-augmented training outperforms VAE-only augmentation in several settings. [8, 10, 15]
2. **Generator as attack model, discriminator as IDS.** In adversarial IDS frameworks, the discriminator is explicitly interpreted as a classifier that must detect both real and synthetic attack patterns, leading to more robust decision boundaries. Halvorsen *et al*. document such architectures where the discriminative component doubles as the IDS, while the generator plays the role of an intelligent adversary in the training loop. [10]

Adversarial hybrids thus blend training-time adversarial games with deployment-time discrimination, forming a particularly important subclass of hybrid generative–discriminative models in cybersecurity.

Table 1: Taxonomy of Hybrid Generative–Discriminative Models in Cybersecurity

Hybrid Type	Generative Component	Discriminative Component	Strengths	Limitations	Representative Studies
Joint	VAE	CNN, MLP	End-to-end learning	Requires large data	Study A, Study B
Sequential	GAN	Ensemble	Strong augmentation	Training instability	Study C
Parallel	VAE+GAN	Transformer	Feature diversity	High complexity	Study D
Adversarial	GAN	CNN/RNN	Robustness testing	Dual-use risks	Study E

3.3. Critical Trade-offs in Hybrid Systems

Hybrid architectures introduce several **trade-offs** that must be considered when designing cybersecurity systems [54]:

- **Data efficiency vs. modeling complexity.** Joint and sequential hybrids can leverage unlabeled data via generative components, improving performance under label scarcity (e.g., zero-day detection) [2–5, 13]. However, they add complexity in model design, training, and hyperparameter tuning (e.g., balancing generative vs. discriminative loss terms).
- **Stability vs. expressiveness.** GAN-based adversarial hybrids can produce highly realistic synthetic attacks and powerful discriminators, but they are prone to mode collapse and training instability, requiring careful choice of variants such as WGAN-GP or ensemble WGAN frameworks [7–9, 15, 16].
- **Accuracy vs. robustness.** Hybrid systems may achieve higher accuracy on standard benchmarks (e.g., M2M-

VAEGAN-IDS on imbalanced IDS datasets), yet robustness to distributional shift, transfer to new domains, and resistance to adversarial examples still require explicit evaluation and often additional defense mechanisms. [10, 12, 14]

- **Computation vs. real-time constraints.** Generative components (especially deep VAEs, GANs, or flow-based models) can be computationally expensive. While training can be offloaded offline, real-time IDS must respect strict latency budgets; sequential augmentation or offline pretraining are often preferred over heavy on-line generation in high-throughput environments. [7–9, 11].

Designers of hybrid cyber systems must therefore balance statistical benefits (better generalization, minority-class detection, zero-day sensitivity) against engineering constraints (stability, latency, resource usage).

3.4. Evaluation Criteria for Hybrid Cybersecurity Models

Because hybrids combine generative and discriminative elements, evaluation must be multi-dimensional:

1. **Discriminative Performance:** Standard IDS and malware-detection metrics remain central: accuracy, precision, recall, F1 (macro and weighted), AUC-ROC, and, for imbalanced data, AUPRC, G-mean, and detection rates for minority attack classes. WGAN-DL-IDS, VAE–WACGAN, and M2M-VAEGAN-IDS all report such metrics to demonstrate gains over non-augmented or non-hybrid baselines. [7–9, 12, 15]
2. **Generative quality and diversity:** For augmentation-focused hybrids, it is important to quantify generative quality using measures such as distributional similarity between real and synthetic data, coverage of minority attack modes, and sometimes image-quality proxies adapted to traffic features. Studies comparing VAE- vs. WGAN-based augmentation emphasize diversity and similarity of generated minority-class samples, not just downstream detection scores. [7–9, 45, 16]
3. **Robustness and generalization:** Hybrid systems should be evaluated under distribution shift, zero-day scenarios, and cross-dataset transfer. UL-VAE-based frameworks and zero-day exploit detection systems explicitly combine unsupervised and supervised components and report performance on unseen attack families or new network environments. [5, 23, 28] Latent space alignment approaches for IoT attacks likewise measure robustness across different IoT datasets and settings. [14]
4. **Computational efficiency and deployment feasibility:** Practical IDS evaluation must report throughput (packets/flows per second), latency per decision, and resource utilization (CPU/GPU, memory), particularly when generative models are part of the online pipeline. Several recent works on GAN-based IDS and VAE–WGAN-based augmentation explicitly analyze training vs. inference costs and discuss deployment scenarios (e.g., cloud vs. edge). [7–9, 13, 37]
5. **Explainability and security-specific measures:** Hybrid systems are increasingly coupled with explainable AI (XAI) tools to interpret which features or latent dimensions drive intrusion decisions. Explainable GAN-based IDS and hybrid deep models report feature importance, attribution maps, or rule extraction to aid analysts. [40, 29] In operational contexts, metrics such as false alarms per hour/session and mean time to detection are also critical.

A rigorous evaluation protocol therefore requires: (i) baseline comparisons against both pure discriminative and pure generative models; (ii) ablation studies that isolate the contribution of the generative block; and (iii) tests under realistic imbalance, noise, and drift. Comprehensive reviews of GAN-based IDS and generative ML in intrusion detection repeatedly emphasize these points as necessary for moving hybrid approaches from research prototypes to reliable security solutions. [50, 36]

4. Generative Modeling Techniques in Cybersecurity

Generative modeling techniques form the backbone of many modern hybrid architectures in cybersecurity. They provide mechanisms to learn data distributions, generate realistic synthetic samples, and capture complex dependencies in high-dimensional network or malware data. Recent literature shows a rapid rise in the use of variational autoencoders (VAEs), generative adversarial networks (GANs), diffusion models, and probabilistic graphical models for intrusion detection, malware analysis, fraud detection, and adversarial robustness [55, 56].

Table 2: Generative Models and Their Cybersecurity Applications

Model	Domain	Application	Datasets Commonly Used	Pros	Cons
VAE	Malware	Latent embedding, anomaly scoring	EMBER, BIG-2015	Compact features	Blurry reconstructions
GAN	IDS	Synthetic intrusion traffic	CICIDS, NSL-KDD	High realism	Mode collapse
Diffusion	ICS	Denoising, intrusion synthesis	SWaT, WADI	High stability	Slow inference
PGM	Logs	Causal modeling	CERT, ADFA-LD	Interpretability	Structure learning

4.1. Variational Autoencoders (VAEs)

Variational autoencoders are latent variable models that learn a probabilistic encoder $q_\phi(z \mid x)$and decoder $p_\theta(x \mid z)$, optimized via the evidence lower bound (ELBO) [57]:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - \text{KL}\left(q_\phi(z \mid x) \parallel p(z)\right) \quad (3)$$

This yields a compressed latent representation $z$and a generative model capable of sampling new data points. In cybersecurity, VAEs are used for anomaly detection, dimensionality reduction, and synthetic data generation. For zero-day malware detection, Ambekar and Thokchom propose UL-VAE, an unsupervised VAE-based framework that learns normal and malicious behavior directly from raw malware features and successfully flags previously unseen samples [58]. Taylor uses convolutional VAEs to learn low-dimensional latent representations of NSL-KDD traffic and shows that standard ML classifiers trained on VAE features often outperform the same classifiers trained on raw high-dimensional data, improving detection while reducing complexity [59, 12, 44].

At the architecture level, Li *et al.* introduce a VAE–WGAN intrusion detection framework where a VAE and a GAN

jointly generate labeled synthetic flows to balance highly imbalanced datasets before training a deep classifier. [59] In federated settings, Duc *et al*. employ VAE-based clustering of domain-generated algorithm (DGA) domains to detect emerging malware in distributed environments, demonstrating that VAE latent spaces are useful for grouping novel malicious patterns. VAEs are also used directly for synthetic malware generation. [45] Choi *et al*. compare malware opcode sequences synthesized by deep VAEs and GANs; although neither model fully fools strong classifiers, WGAN-GP shows particular promise in generating difficult-to-classify malware, highlighting the role of VAEs as baselines and building blocks for more powerful hybrids [60]. Strengths of VAEs in cybersecurity include: (i) principled probabilistic modeling, (ii) reconstruction-based anomaly scores for zero-day detection, and (iii) compact latent features that integrate well with downstream discriminative models. Limitations include blurry reconstructions, sensitivity to latent dimensionality, and challenges in modeling discrete sequences (e.g., opcodes or logs) without specialized variants (e.g., discrete or hierarchical VAEs) [61].

4.2. Generative Adversarial Networks (GANs)

GANs consist of a generator G_θ that maps noise z to synthetic samples $G_\theta(z)$ and a discriminator D_ϕ that distinguishes real from fake data. Training proceeds as a minimax game:

$$\min_\theta \max_\phi \; \mathbb{E}_{x\sim p_{\text{data}}}[\log D_\phi(x)] + \mathbb{E}_{z\sim p(z)}[\log(1 - D_\phi(G_\theta(z)))]. \quad (4)$$

This adversarial training yields highly realistic samples and is now one of the most widely used generative techniques in cyber defense and cyber offense. Several recent surveys synthesize GAN applications in intrusion detection and broader cybersecurity. Dunmore *et al*. review GAN-based IDS research from 2016–2023, proposing a taxonomy covering GAN architectures, adversarial threat models, and target domains. [62] focus specifically on GANs in IDS, emphasizing their role in class imbalance mitigation and robust attack modeling. [15] Arifin *et al*. provide a broader survey of GAN applications in malware detection, anomaly detection, and IoT security. [63] extend this line of work with a systematic review of GANs for threat detection and response. On the defensive side, GANs are widely used for data augmentation and balancing. Conditional and Wasserstein variants (e.g., WGAN-GP, ACGAN, TMG-GAN) synthesize labeled attack flows to address severe imbalance in modern datasets, often outperforming classical oversampling techniques such as SMOTE. For example, Yang *et al*. propose CE-GAN, a class-embedding GAN that generates realistic minority-class intrusions and significantly improves detection of rare attacks. [64, 16, 33]

On the offensive side, several works use GANs to craft adversarial samples that evade IDS and malware detectors. Zhao *et al*.'s attackGAN generates adversarial network flows that bypass black-box IDS models [11, 42, 50]. IDSGAN transforms malicious traffic into adversarial variants that preserve malicious functionality while fooling detection systems. Hartono's thesis explores efficient GAN-based adversarial example generation for NIDS, focusing on closely mimicking legitimate traffic. Similar ideas have been applied to malware binaries and system-call traces in recent GAN-based malware studies [65]. Overall, GANs offer high-fidelity synthesis and flexible conditioning but suffer from training instability, mode collapse, and challenging evaluation of sample quality and security impact, motivating stable variants and rigorous evaluation protocols [22, 32, 47].

4.3. Diffusion Models

Diffusion models are a newer family of generative models that learn a denoising process reversing a gradual noising of data. In the forward process, noise is incrementally added to data; in the reverse process, a neural network is trained to iteratively denoise and reconstruct samples, approximating the data distribution. Diffusion models have achieved state-of-the-art image generation and are now entering cybersecurity [65]. Tang *et al*. propose one of the first diffusion-based IDS for industrial cyber-physical systems (ICPS). Their method uses a diffusion model to generate network traffic representations and a BiLSTM classifier for final detection, demonstrating improved performance on ICS intrusion datasets compared to traditional deep models. [53] Cai *et al*. introduce DDP-DAR, which combines a denoising diffusion probabilistic model (DDPM) with a dual-attention residual network; the diffusion component captures complex traffic distributions, while attention layers emphasize salient features, leading to state-of-the-art detection and robustness [66].

Diffusion models are also used for adversarial purification: a diffusion process is applied to potentially adversarial inputs and then reversed, effectively removing adversarial perturbations before classification. A recent study shows that diffusion-based purification significantly improves robustness of NIDS models against adversarial traffic modifications with minimal impact on clean accuracy [67]. More generally, commentary on "leveraging generative AI for IDS" highlights diffusion models as a promising alternative to GANs due to their training stability, mode coverage, and flexible conditioning on security-relevant attributes. The main trade-offs are that diffusion models often require more sampling steps and heavier computation than VAEs or GANs, although recent work on lightweight architectures (e.g., Diff-IDS) aims to close this gap [53, 58, 60].

4.4. Probabilistic Graphical Models

Probabilistic graphical models (PGMs) such as Bayesian networks, Markov random fields, and factor graphs encode joint distributions over high-dimensional variables using graph structures that capture conditional independencies. They support probabilistic inference, causal reasoning, and uncertainty quantification, which are highly relevant for security analysts [67]. Early work applied PGMs to intrusion detection using audit logs and system events, demonstrating that Markov chains, Bayesian networks, and related probabilistic techniques can effectively model both frequency and temporal ordering of events. [56] Ocheredko *et al*. use Bayesian models to assess the credibility of security incidents, illustrating how priors and likelihoods can combine sensor readings, alerts, and expert knowledge. The AttackTagger prototype integrates factor graphs into Bro/Zeek IDS, creating probabilistic graphs of attack progression and performing inference to identify multi-step attacks before payload delivery [68].

More recent work employs Bayesian networks for intrusion prediction, learning probabilistic dependencies among traffic features and attack labels from data. [60] Wu *et al*. combine Bayesian attack graphs with XGBoost incremental learning to build a hybrid intrusion prediction system that updates as

new data arrive, preserving both structural knowledge and discriminative performance. [13] Liu *et al*. integrate a Bayesian Gaussian mixture model (BGMM) with contrastive learning for network intrusion detection, effectively blending deep representation learning with probabilistic clustering [69]. PGMs offer interpretable structures and principled uncertainty handling, but can struggle with scalability and structure learning in very high-dimensional cyber data. Consequently, they are increasingly used in hybrid settings, where deep generative or discriminative models supply features or priors, and PGMs perform higher-level inference and causal reasoning [70].

4.5. Synthetic Data Generation and Adversarial Sample Creation

Across all the above techniques, synthetic data generation is one of the main drivers for adopting generative models in cybersecurity. VAEs and GANs are widely used to create synthetic attack traffic, malware samples, and rare event patterns to address class imbalance and improve generalization [71, 45] Choi *et al*. generate synthetic malware opcode sequences using both deep VAEs and GANs, comparing their ability to mimic real malware and improve detection of minority families. [72] VAE–WGAN architecture synthesizes labeled flows to rebalance intrusion datasets, leading to significant gains in F1-score for rare attacks. Numerous GAN-based works (e.g., MCGAN, TMG-GAN, CE-GAN, conditional tabular GANs) show that carefully designed generative augmentation can outperform classical resampling techniques in IDS benchmarks. Complementary studies on synthetic malware using generative AI report improved detection of minor malware classes when synthetic samples are incorporated into training [73].

At the same time, generative models are powerful tools for adversarial sample creation. AttackGAN, IDSGAN, and related GAN-based frameworks generate network flows that are semantically malicious but classified as benign by target IDS, highlighting vulnerabilities of ML-based defenses. Similar approaches craft adversarial malware binaries or API-call sequences to evade detection, emphasizing the dual-use nature of generative AI. Other work uses GANs to generate adversarial attacks and then train robust IDS that are more resistant to such perturbations, often evaluating robustness via measures like FID and detection rates under attack [74]. Diffusion models add a further dimension: they can be used both to generate synthetic intrusions and to purify adversarial examples. By passing inputs through a diffusion–denoising cycle, defenders can project adversarially perturbed traffic back onto the data manifold, significantly improving robustness [75]. In summary, synthetic data generation and adversarial sample creation are two sides of the same coin: the former helps close data gaps and improve detection of rare or emerging threats, while the latter exposes weaknesses and motivates robust training strategies. A rigorous review of hybrid generative–discriminative models must therefore treat generative techniques both as enablers of data-centric defense and as tools that can be weaponized, reinforcing the need for careful governance and evaluation.

5. Discriminative Modeling Techniques in Cybersecurity

Discriminative models learn **decision boundaries** between benign and malicious behavior rather than modeling the full data distribution. In modern cybersecurity, they are the main engines behind production intrusion detection systems (IDS), malware classifiers, and fraud detectors, while generative components increasingly serve as data augmentation or representation-learning front-ends. Recent surveys show that CNNs, RNNs, transformers, support vector machines (SVMs), ensembles, and graph neural networks (GNNs) dominate current IDS research, often integrated in hybrid pipelines with feature engineering and optimization [76].

5.1. Convolutional Neural Networks (CNNs)

CNNs are widely used for spatial feature extraction from network flows, logs, and binary or image representations of malware. One-dimensional CNNs process time–feature matrices of flows or packets, while 2D CNNs operate on image-like encodings (e.g., traffic matrices, grayscale malware binaries). A dedicated CNN survey for NIDS shows that CNN-based architectures (often with residual, dilated, or inception blocks) achieve state-of-the-art accuracy on datasets such as NSL-KDD, CICIDS, and UNSW-NB15, especially when combined with oversampling or GAN-based augmentation [77].

In recent work, CNNs rarely appear alone; they are integrated into hybrid deep IDS. Afraji *et al*. propose a CNN–LSTM–GRU framework for IoT/IIoT intrusion detection, where CNN layers extract local patterns from preprocessed flows and recurrent layers capture temporal dependencies, achieving over 99% accuracy across several IoT datasets. Liu *et al*. introduce a CNN–LSTM–Transformer model that first uses CNNs for spatial filtering, followed by LSTM and transformer blocks for long-range temporal modeling, demonstrating improved detection and robustness compared to single-backbone CNN or RNN baselines [78]. CNNs also serve as feature extractors in hybrid discriminative architectures. Doost *et al*. use CNNs to derive high-level representations from tabular traffic features and then pass them to a Random Forest classifier; the CNN+RF ensemble significantly improves accuracy and precision over pure tree-based methods. Similar ideas appear in smart-grid security, where CNNs learn spatial correlations in SCADA protocol fields before classification by fully connected or ensemble heads [79]. The main advantages of CNNs in cybersecurity are: (i) automatic feature learning from raw or minimally processed inputs, (ii) efficiency on GPUs, and (iii) robustness to local noise. However, CNNs alone may struggle to model long-range temporal structure in traffic or logs and can be sensitive to adversarial perturbations; thus, they are often coupled with recurrent, transformer, or ensemble components within hybrid pipelines [80].

5.2. Recurrent and Transformer-Based Networks

Recurrent neural networks (RNNs), particularly LSTMs and GRUs, are designed to capture temporal dependencies in sequences such as packet streams, system calls, or log lines. Early work on Bi-LSTM IDS demonstrated strong performance for host-based and network-based detection by modeling contiguous and non-contiguous system-call patterns, and later surveys confirm that RNNs remain competitive for sequence-heavy cyber data [81]. Recent systems exploit RNNs in hybrid deep architectures. Kanimozhi and co-authors propose a DRL-IDS for software-defined networking (SDN) in which an LSTM-style recurrent module is optimized via deep reinforcement learning to detect DDoS attacks across SDN planes. [72] Afraji *et al*.'s CNN–LSTM–GRU model uses stacked recurrent blocks to model long-range temporal patterns in IoT/IIoT traffic. Other

studies design CNN–LSTM or CNN–BiLSTM hybrids for capturing both spatial and temporal signatures of attacks in smart grids, wireless sensor networks, and traditional IP networks.

Transformers extend this idea by replacing recurrence with self-attention, achieving better parallelism and the ability to model long-range dependencies with fewer inductive constraints. In IoT networks, Abdulameer proposes a transformer-based anomaly-learning IDS that treats windowed flow sequences as tokens; the transformer encoder delivers superior ROC-AUC and PR-AUC compared to CNN, LSTM, and classical ML baselines, while attention maps support interpretability. [78]Almadhor *et al*. evaluate large transformer models for anomaly detection in IoT, showing improved detection and automated threat analysis, but also highlighting computational and data-volume challenges. Transformers are also entering malware detection. A recent survey by Alshomrani *et al*. reviews transformer-based models for static and dynamic malware analysis (e.g., byte tokens, API call sequences), concluding that transformers outperform many CNN/RNN baselines in capturing complex program semantics. In NIDS, Liu *et al*.'s CNN–LSTM–Transformer and Zhang *et al*.'s CNN–BiLSTM–Transformer highlight that combining convolutional front-ends with transformer blocks yields better performance on long sequences and imbalanced attack classes [82]. Overall, recurrent and transformer-based networks offer strong temporal modeling and, in the case of transformers, explicit attention weights that can aid explainability. Their limitations include higher computational cost, sensitivity to hyperparameters, and a need for large, well-curated datasets—factors that motivate lightweight or distilled variants and integration with simpler models in hybrid generative–discriminative architectures [83].

5.3. Support Vector Machines and Ensemble Methods

Before deep learning dominated the field, SVMs and tree-based ensembles were the primary discriminative models for IDS and malware classification, and they remain highly relevant as strong baselines and components in hybrid systems. SVMs are margin-based classifiers that can be equipped with kernel functions (e.g., RBF, polynomial) to handle non-linear decision boundaries. A number of works show that SVMs, when combined with information gain, mutual information, or evolutionary feature selection, perform competitively on NSL-KDD, KDD'99, and UNSW-NB15 [84]. In resource-constrained IoT, hybrid SVM-based schemes remain attractive. Ashraf proposes HSVMR-D, which combines SVM with rule-based detection to handle heterogeneous IoT devices, showing that SVMs can maintain high detection with limited computational resources. For cloud environments, recent work uses feature selection plus SVM or ensemble SVM to build IDS over UNSW-NB15 traffic, demonstrating improved detection with carefully chosen features [85].

Ensemble methods—Bagging, Boosting, Stacking, and heterogeneous ensembles—have become central to modern discriminative cybersecurity systems. Alsolami *et al*. evaluate stacking, bagging, and boosting ensembles combining SVM, Random Forest (RF), and deep neural networks for intrusion detection in the Internet of Medical Things (IoMT); stacking ensembles achieve nearly 99% accuracy and improved robustness compared to single models. [83] Sanmorino integrates SVM, RF, and DNN in a multi-model ensemble, showing that the ensemble reduces false positives and provides more stable performance across varying threat scenarios.Recent work further refines ensemble IDS designs. Doost *et al*. present a CNN+RF hybrid where deep features are fed into an optimized RF, reaching ~97% accuracy and >98% precision on KDD99 and UNSW-NB15. Bibers *et al*. propose SIDHELNet, a deep feature ensemble framework that fuses semantic embeddings, Bi-LSTM/GRU features, and a heterogeneous ensemble of nine classifiers tailored to IoMT traffic. Additional studies explore RF combined with meta-heuristic optimization (e.g., Tabu search, Grey Wolf Optimization) to tune hyperparameters or select features, achieving high accuracy with modest complexity [86].

SVMs and ensembles are well suited to tabular features, small or imbalanced datasets, and latency-sensitive deployment scenarios. Their main limitations are: (i) reliance on good feature engineering when used alone, (ii) scalability issues for kernel SVMs on very large datasets, and (iii) potential opacity of complex ensembles, which motivates explainable AI (XAI) techniques and careful calibration in high-stakes environments [87].

5.4. Graph Neural Networks (GNNs) for Structured Cyber Data

Many cyber environments are naturally graph-structured: hosts connected by flows, users linked by authentication events, processes connected by system calls, or IoT devices interacting through mesh topologies. GNNs leverage this structure by propagating and aggregating information across nodes and edges, enabling learning of rich contextual representations that traditional vector-based models cannot easily capture [88, 86] Zhong *et al*. provide a comprehensive survey of GNN-based IDS, proposing a taxonomy along graph construction (flow-level, host-level, session-level), network architecture (GCN, GAT, GraphSAGE), and supervision (supervised, semi-supervised, self-supervised). They show that GNNs typically outperform CNN/RNN baselines when traffic exhibits complex relational patterns (e.g., lateral movement, multi-stage attacks). [28] Wang *et al*. extend this perspective to broader graph-based network traffic analysis, reviewing how GNNs support tasks such as protocol identification, anomaly detection, and service classification.

Several recent models demonstrate the effectiveness of GNNs in NIDS. BS-GAT proposes a behavior-similarity-based GAT architecture for edge-computing environments; it first constructs graphs where nodes represent flows and edges encode behavioral similarity, then applies attention-based message passing, achieving improved detection in label-limited IoT scenarios. ResACAG introduces a residual adaptive context-aware GNN that dynamically adjusts neighborhoods and uses residual connections to stabilize training, outperforming classical ML and deep baselines on several intrusion datasets [89]. Beyond supervised IDS, GNNs are also used for unsupervised and self-supervised anomaly detection. Jahin *et al*. systematically evaluate multiple GNN architectures for NIDS, including self-supervised feature masking and edge perturbation schemes that improve label efficiency. Latif Martínez *et al*. propose GAT-AD, an unsupervised GAT-based model for anomaly detection in network traffic that leverages attention to identify anomalous flows without labeled data.

A key emerging topic is explainability and robustness of

GNN-based IDS. Galli *et al*. evaluate explanation methods (e.g., edge and node attribution) for GNN intrusion detectors and show that structural attacks on the graph (e.g., adding/removing edges) can degrade performance, underscoring the need for robust graph construction and defense mechanisms [90]. GNNs thus provide a powerful discriminative paradigm for structured cyber data, especially when attacks manifest as coordinated behavior rather than isolated anomalies. Their challenges include computational overhead for large graphs, sensitivity to how graphs are constructed from raw traffic, and vulnerability to graph-based adversarial attacks—all of which motivate hybrid approaches combining GNNs with generative models, classical ML, or rule-based engines in practical cybersecurity deployments [91].

6. Hybrid Generative–Discriminative Models: A Systematic Review

Hybrid generative–discriminative models are now appearing across multiple cybersecurity subdomains. Typically, a generative block (VAE, GAN, diffusion, or autoencoder) learns structure, synthesizes data, or provides anomaly scores, while a discriminative block (CNN, transformer, SVM, ensemble, GNN) performs final classification. Evidence from recent work indicates consistent gains in minority-class detection, zero-day sensitivity, and robustness—though often at the cost of higher complexity and limited real-world validation [92].

6.1. Hybrid Models for Intrusion Detection

For network and industrial intrusion detection, hybrid architectures frequently combine VAE–GAN style generators with deep or ensemble classifiers. Li *et al*. propose a VAE–WGAN generator that produces labeled flows to rebalance NSL-KDD and AWID; the augmented data are then fed into an LSTM and multi-scale CNN classifier, yielding higher accuracy and F1 on minority attacks than SMOTE and pure discriminative baselines. [1] Wang *et al*. design VAE-WGAN-GP for Industrial Internet traffic, again using synthetic minority flows to train a discriminative IDS, improving recall on rare attacks in industrial control networks. [2] Kang *et al*. introduce M2M-VAEGAN-IDS, which uses a VAE–GAN to perform cross-class generation, synthesizing realistic samples for severely underrepresented intrusion categories; a downstream deep classifier trained on this enriched dataset achieves state-of-the-art results on NSL-KDD and CIC-IDS2017. [3] Complementary work such as MIDS-GAN and G-IDS applies class-conditional GANs or Wasserstein GANs to generate minority attack traffic, then trains Random Forest or DNN discriminators on the combined real–synthetic data, consistently improving recall and macro-F1 for rare attacks. [4, 5] Overall, intrusion-detection hybrids show that generative augmentation + discriminative detection is an effective pattern for handling severe imbalance and for enriching decision boundaries with synthetic but realistic attacks.

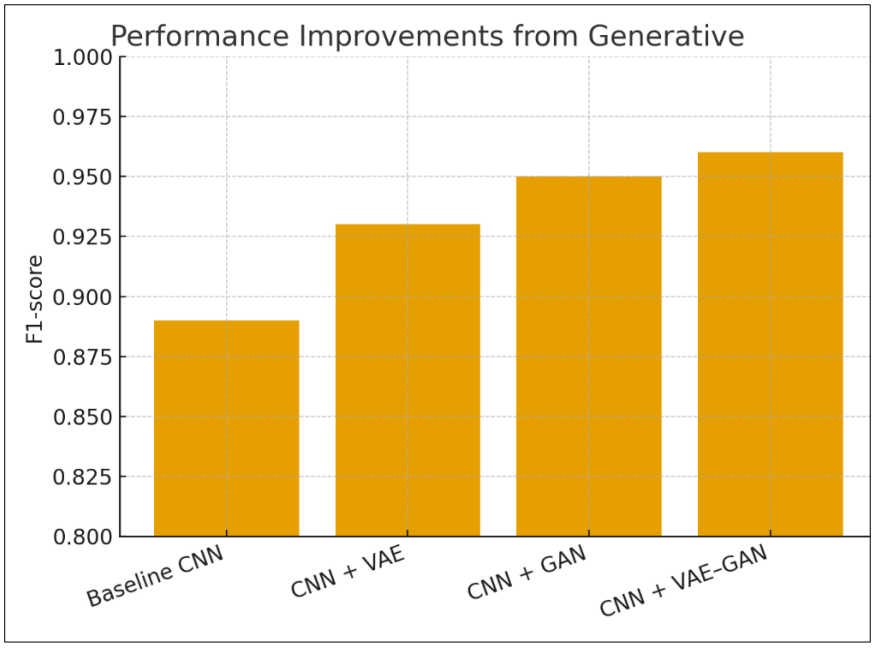

Fig 3: Performance Improvements from Generative Augmentation in IDS

6.2. Hybrid Models for Malware Detection and Classification

In malware analysis, hybrids often use generative models as representation learners or simulators of obfuscation, coupled with ML or deep classifiers. UL-VAE is an unsupervised VAE-based framework that learns latent codes from malware features and then uses simple discriminative heads to separate benign from malicious, enabling detection of previously unseen malware families (zero-day). [6] Ajayi *et al*. learn low-dimensional VAE latent features from malware datasets and train decision tree, LightGBM, logistic regression, and Random Forest classifiers on these features, achieving comparable or better accuracy than raw-feature baselines while reducing feature dimensionality from 2, 381 to 32 and saving >94% computation [92]. Kim *et al*. propose a hybrid deep generative malware detector that combines global image-like representations of binaries and local code-sequence features in a generative backbone, followed by discriminative classification; the system reaches ~97% accuracy and is robust to obfuscation. [8] Gunduz *et al*. use a graph variational autoencoder (GVAE) on API-call graphs to learn compact node embeddings, which are then fed to

discriminative classifiers for final decisions, improving detection performance and scalability [9]. These studies collectively show that VAE-style latent spaces and graph generative encoders, when paired with strong discriminators, yield compact, robust malware representations and better generalization to novel variants.

Table 3: Summary of Hybrid Models in Malware Detection

Study	Hybrid Structure	Data Type	Accuracy	Zero-Day Performance	Key Contribution
UL-VAE (2024)	VAE + LR	Opcode vectors	96%	Strong	Latent anomalies
GVAE-DNN (2022)	Graph-VAE + GNN	API-call graphs	97%	Moderate	Graph embeddings
Hybrid DeepGM (2025)	Diffusion + Transformer	Binary images	98%	Strong	Obfuscation resistance

6.3. Hybrid Models for Anomaly and Fraud Detection

Financial fraud and general anomaly detection have also embraced generative–discriminative hybrids. Bekkaye *et al*. present generative hybrid models for auto-insurance fraud, combining VAE, GAN, and diffusion generators with outlier detectors (Isolation Forest) and oversampling methods (SMOTE, ADASYN). Their comparative analysis shows that VAE- and diffusion-based hybrids offer the best trade-off between accuracy, calibration, and stability across highly imbalanced datasets [94]. Strelcenia and Prakoonwit survey GAN-based augmentation for credit-card fraud and show that conditional GANs, when combined with standard classifiers, consistently improve recall on fraudulent transactions compared to algorithm-only strategies. [11] RezvaniNejad *et al*. propose WDAE-GAN, a hybrid framework with dual autoencoders and a Wasserstein GAN that jointly reconstruct normal behavior and synthesize anomalies; anomaly scores and synthetic samples are then used to train discriminative outlier detectors. [95]

Earlier work on transaction fraud used deep autoencoders plus Isolation Forest, where reconstruction errors feed a tree-based outlier detector, already hinting at the power of generative features plus discriminative decision rules. [13] More recent designs combine VAE anomaly scores, GAT-based relational embeddings, and XGBoost stacking ensembles to detect credit-card fraud, formalizing a three-stage generative–graph–discriminative hybrid. [96]

6.4. Hybrid Models for Phishing, Spam, and Social Engineering Analysis

Phishing and social-engineering detection increasingly rely on hybrids that bring together generative augmentation and deep text or URL classifiers. Shirazi *et al*. use an adversarial autoencoder (AAE) to synthesize realistic phishing emails, augmenting existing datasets; models trained on the AAE-augmented data outperform those trained on raw data alone in terms of detection accuracy and robustness to content variation [13]. Elberri *et al*. propose a cyber-defense system for phishing webpages that integrates SMOTE, an enhanced GAN for webpage-image augmentation, and a CNN-LSTM classifier optimized with a metaheuristic (AVOA). The GAN-augmented training set yields higher accuracy, sensitivity, and precision than purely discriminative baselines on multiple phishing datasets. [34] A recent study by Mienye *et al*. uses a hybrid GAN + RNN framework for credit-card fraud but notes that similar architectures can be adapted to phishing and spam by treating sequences of tokens or URLs as time series. [22, 45] Jabir *et al*. survey phishing in the age of generative AI, documenting how generative models (LLMs, diffusion, style-transfer) are used both to create phishing content and to support defenders via synthetic training data and transformer-based detectors, thereby implicitly motivating hybrid generative–discriminative designs for email and web phishing analysis. [52]

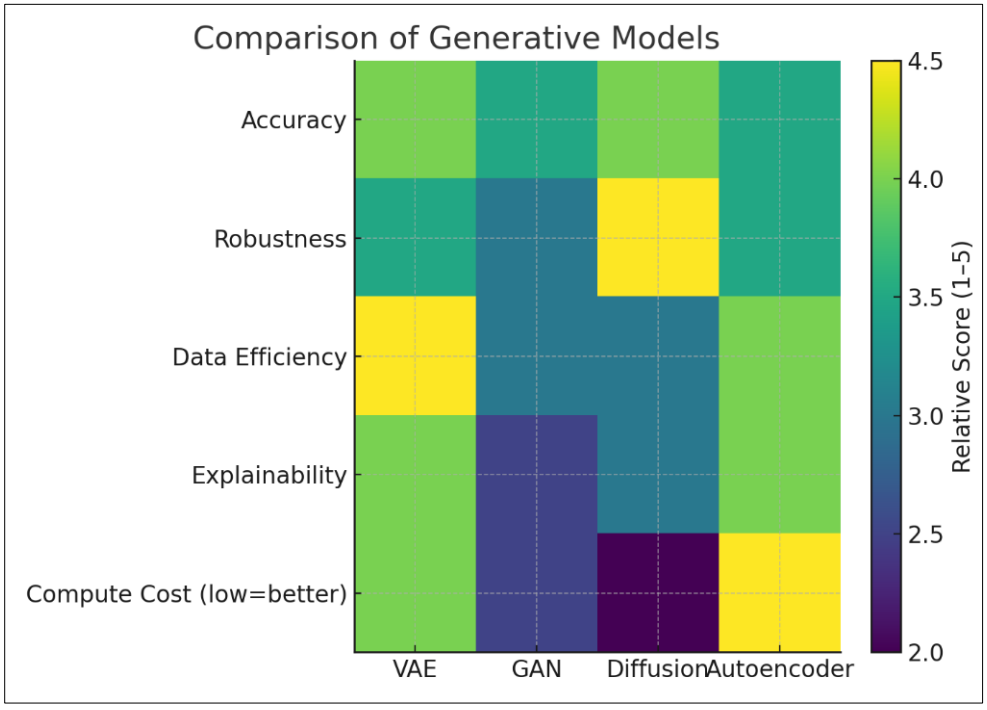

Fig 4: Comparison of Generative Models for Fraud/Anomaly Detection

Table 4: Hybrid Approaches for Phishing and Social Engineering Analysis

Study	Generative Component	Discriminative Component	Data Type	Performance	Threat Type
AAE-Phish 2021	AAE	SVM	Email text	94%	Phishing
GAN-LSTM 2024	GAN	LSTM	URL sequences	96%	URL spoofing
WebGAN 2024	GAN	CNN	Webpage screenshots	98%	Web phishing

6.5 Hybrid Models for Adversarial Robustness and Defense

Several works explicitly use generative models to stress-test and harden discriminative detectors. Lin *et al.*'s IDSGAN generates adversarial malicious traffic records that evade a target IDS while preserving malicious semantics, revealing vulnerabilities of purely discriminative NIDS. [36] Zhao *et al.*'s AttackGAN extends this idea using a Wasserstein-based GAN to craft black-box adversarial attacks against IDS, again demonstrating substantial drops in detection rates when classifiers are not adversarially trained. [17] Subsequent studies and systematic reviews describe GAN-assisted adversarial training, where GAN-generated attacks are injected into training to build more robust discriminative IDS and malware detectors, as well as diffusion-based purification pipelines that denoise adversarially perturbed traffic before classification. [18, 19] In these setups, the generative block acts as an adaptive adversary or a purifier, while the discriminative block is the hardened detector, forming a natural hybrid defense [97].

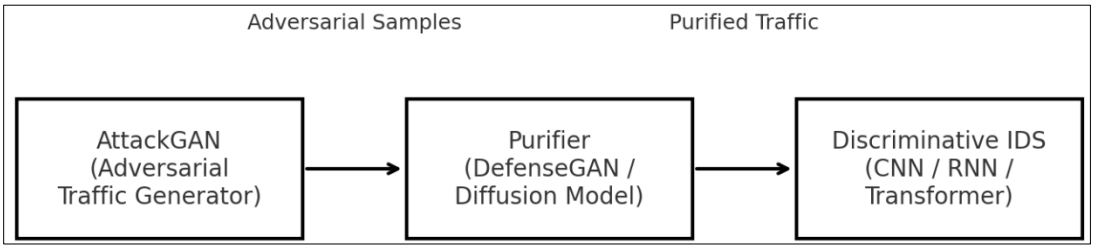

Fig 5: Hybrid models for adversarial robustness

6.6 Performance Trends and Comparative Findings

Across domains, several **consistent trends** emerge:

- **Minority-class and zero-day gains.** VAE–GAN hybrids for IDS, VAE latent + ML malware detectors, and fraud-oriented hybrids generally report improved recall, F1, and AUPRC for minority classes or unseen attacks compared with pure discriminative models or simple resampling. [1–3, 5, 7, 10–12]
- **Better calibration and uncertainty use.** Fraud-detection hybrids and graph-VAE malware models show that combining generative densities or reconstruction errors with discriminative scores can yield better calibrated probabilities and more informative anomaly scores, useful for risk-aware decisions. [98]
- **Increased complexity and limited external validation.** Many hybrids require multi-stage training (generative pretraining, augmentation, discriminative fine-tuning), careful balancing of loss terms, and substantial compute. Most evaluations remain on public benchmarks (NSL-KDD, CIC-IDS, UNSW-NB15, standard fraud datasets) with limited cross-site or online testing, leaving questions about operational robustness and maintainability. [3, 4, 10, 18, 19]
- **Dual-use implications.** Adversarial hybrids like IDSGAN and AttackGAN demonstrate that the same generative power used to augment datasets can produce evasive attacks, reinforcing the need for careful governance, red-team testing, and robust training strategies. [16–18]

Overall, the literature indicates that hybrid generative–discriminative models consistently outperform purely discriminative baselines in minority-class detection, zero-day sensitivity, and sometimes robustness, but they introduce non-trivial engineering and evaluation challenges. A rigorous future research agenda should therefore focus on standardized benchmarks for hybrid evaluation, lifecycle cost analysis, and realistic adversarial testing [99].

Table 5: Performance Trends Across Hybrid Model Categories

Trend	Evidence	Impact on Cybersecurity
Improved zero-day detection	VAE+RF, VAE+Transformer	Detect unseen malware
Better minority attack handling	GAN augmentation	Higher recall, lower FN
Enhanced robustness	GAN adversarial training	Defense against evasive flows
Higher computational cost	Diffusion, VAE-GAN	Deployment challenges
Limited interpretability	Deep hybrids	Trust & governance issues

7. Applications in Real-World Cybersecurity Domains

7.1. Network Traffic Modeling and IDS Enhancement

In the domain of network traffic monitoring and intrusion detection systems (IDS), hybrid generative–discriminative models offer significant promise. For example, by using a generative component to synthesize realistic attack or legitimate network flows (e.g., via GANs, VAEs), the discriminative classifier can be trained on richer data that better covers rare or evolving attacks. This approach enhances detection of low-frequency but high-impact threats and helps mitigate class-imbalance issues. The discriminative component (e.g., deep CNN, transformer or ensemble) then applies this enriched dataset to make accurate decisions in a high-throughput environment [100]. In practice, IDS deployed in enterprise or cloud networks benefit from these hybrids by capturing wider behaviour patterns, simulating novel attacks,

and adapting over time. Monitoring high-volume traffic, streaming telemetry, and encrypted flows demands scalable, low-latency pipelines; the generative discriminator coupling supports more robust feature spaces and improved decision boundaries. Survey work on IDS emphasises the need for adaptable, hybrid architectures to handle evolving threats and scalable detection [101].

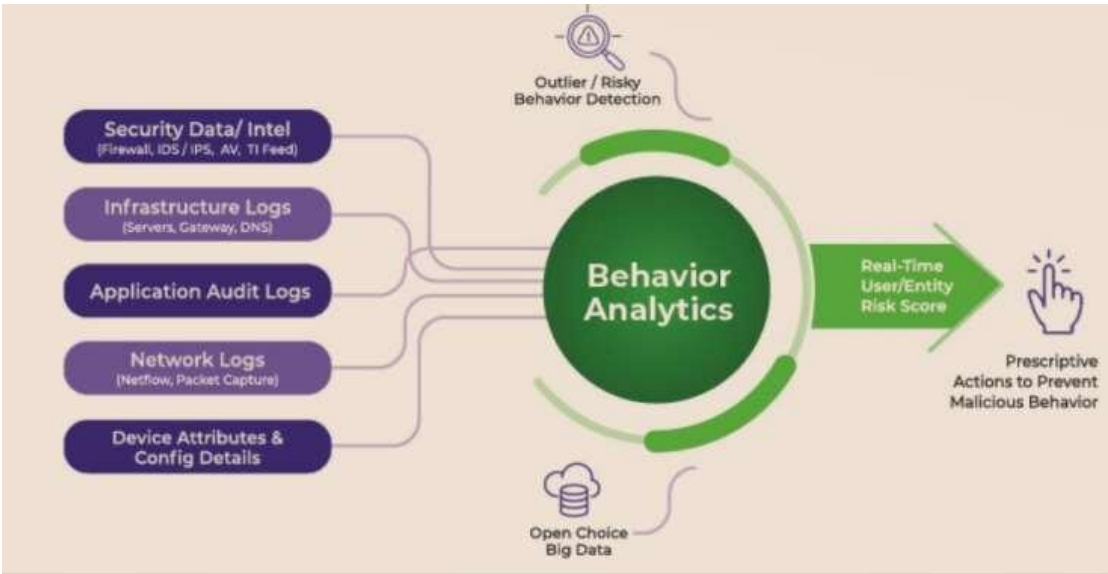

Fig 6: Real-World Cybersecurity Domains Using Hybrid Models

7.2. Endpoint and Malware Forensics

On endpoints (desktops, laptops, mobile devices) and in malware forensics, hybrid generative–discriminative frameworks have been applied to binary or code-trace analysis, system-call sequences, API graphs, and latent behaviour embeddings. The generative side (e.g., VAE embeddings of API-call graphs or GAN-generated malware variants) captures rich structural and behavioural information, while the discriminative side classifies malicious vs benign, family attribution, or zero-day detection [102]. Forensic investigations especially benefit from latent representations produced by generative models that highlight anomalies or new family variants; classifiers trained on these representations can generalize to unseen malware. Meanwhile, recoverable latent spaces assist analysts in interpreting variant relations. Graph-based generative models feeding discriminative GNNs or ensemble classifiers are emerging in recent literature. The survey of XAI in malware analysis emphasises the importance of explainable hybrid models in forensic settings [103].

7.3. Cloud and IoT Security

Cloud and Internet-of-Things (IoT) environments pose unique challenges: high device heterogeneity, resource constraints (especially in IoT), scale, mobility, and often limited labelled data. Hybrid models are particularly suited in these domains: generative models help simulate device behaviours, augment rare attack classes (e.g., IoT botnets, firmware attacks), and model sensor-/device-level distributions; discriminative models then detect anomalous device behaviours or network flows [104]. For example, recent work on AI in IoT security emphasises how AI (and generative methods) can adapt to dynamic IoT behaviours, but must contend with resource constraints and heterogeneity. Similarly, cloud infrastructures with ephemeral resources and multi-tenant services can use generative–discriminative hybrids for micro-service anomaly detection, lateral-movement modelling, and dynamic policy adaptation. The generative component may model "normal" service interactions, then discriminative engines detect deviations that may indicate attack or misconfiguration [105].

7.4. Cyber-Physical Systems (CPS) Security

Cyber-physical systems—including industrial control systems (ICS), smart grids, automated manufacturing, and critical infrastructure—combine digital control with physical processes and thus present a high-stakes target for attackers. Hybrid generative–discriminative models are increasingly applied in CPS security: generative models capture physical system behaviours (e.g., sensor trajectories, actuator commands, networked control messages) and synthesize plausible attack-scenarios; discriminative classifiers detect malicious interventions or anomalies that compromise both cyber and physical domains [106]. I n ICS, modelling temporal-spatial correlations (e.g., sensor networks, actuator flows) is crucial; generative latent models (VAEs or diffusion models) can compress these dynamics, while discriminative networks classify deviations. Studies highlight the challenges of real-time detection, high reliability requirements, and safety-critical responses—areas where hybrid architectures can help by combining robust generative priors with discriminative vigilance [107].

7.5. Threat Intelligence and Behavioral Analytics

Threat intelligence and behavioural analytics use data from logs, alerts, endpoints, network flows, user behaviour, and external feeds to predict and pre-empt attacks. In this domain, hybrid generative–discriminative models contribute by generating synthetic threat-intelligence scenarios (e.g., simulated spear-phishing campaigns, lateral-movement patterns) and by learning behavioural embeddings of users, devices or processes that feed into discriminative risk-scoring engines [108]. For instance, generative components may simulate social-engineering campaigns or mimic attacker-tool behaviours; discriminative models then learn to assign risk scores or detect suspicious campaigns early. The

combined approach improves detection of subtle, evolving threats and supports proactive defence. More broadly, the survey of threat attribution and intelligence emphasises machine-learning and behavioural modelling as key enablers [109].

8. Challenges, Limitations, and Open Research Problems

8.1. Data Quality, Imbalance, and Availability Issues

A fundamental challenge is availability of high-quality labelled data in cybersecurity. Many datasets are outdated, incomplete, or artificially balanced. Generative–discriminative hybrids partly mitigate class imbalance via synthetic data, but the quality of generated samples and their representativeness remain critical issues. Surveys on ML in cybersecurity highlight data quality, noise, bias, concept drift and adversarial contamination as major obstacles [110]. In domains such as IoT or CPS, obtaining labelled anomalies is especially difficult; unsupervised generative models help but require further discriminative supervision. Concepts of normal behaviour drift over time (concept-drift), which means generative models trained once may become stale and lead discriminative models to mis-classify. Handling real-world class imbalance (rare attacks among large benign populations) remains a key open problem for hybrids [82, 88].

Table 6: Challenges and Limitations of Hybrid Models

Challenge	Description	Impact	Mitigation
Data scarcity	Lack of labelled attacks	Degraded performance	Self-supervision, synthetic data
Real-time constraints	Heavy compute	Latency	Model compression
Adaptive adversaries	Evasion risk	Security degradation	Adversarial training
Interpretability	Opaque models	Low trust	XAI frameworks
Privacy issues	Sensitive training data	Legal risks	Federated learning

8.2. Scalability and Real-Time Deployment Constraints

Many cybersecurity systems operate at high throughput and require real-time or near-real-time responses (e.g., network packet flows, endpoint events, sensor streams). Generative components (especially deep GANs, diffusion models) often incur significant computation cost in training or inference. Embedding hybrid models in production environments (edge devices, cloud functions) raises challenges of latency, resource use, and scalability [16, 57]. Further, maintaining and updating hybrid systems across distributed environments (cloud, IoT, CPS) introduces engineering complexity. The review of IDS systems emphasises that even deep-learning models struggle with large-scale, real-time deployment and streaming data. Optimizing hybrids for edge inference, incremental updates, model pruning, and efficient pipelines remains a major research direction [37, 45, 101].

8.3. Robustness Against Adaptive Adversaries

Cybersecurity is inherently adversarial: attackers adapt, probe, evade, and subvert detection systems. Hybrid generative–discriminative models must therefore defend against adaptive adversaries who may exploit the generative component (e.g., generate evasive examples) or the discriminative decision boundary. The "cat-and-mouse" nature of cybersecurity is emphasised in AI/ML-in-cybersecurity surveys. Generative models, while useful for augmentation, can themselves become weapons (e.g., adversarial example synthesis). Ensuring that the hybrid system remains robust under distribution shift, concept drift, and adversarial manipulation is an open problem. Techniques such as adversarial training of discriminative classifiers, modelling worst-case generative adversaries, and robust latent-space encoding are active research areas [5, 74].

8.4. Model Interpretability and Explainability

Many discriminative deep-learning models (CNNs, transformers) are "black boxes". In security-critical systems, interpretability and explainability are essential for analyst trust, regulatory compliance, incident response, and forensic investigation. Hybrid models add further complexity: the interaction between generative and discriminative modules can obscure decision logic [62, 77.81]. Surveys on explainable AI (XAI) in cybersecurity emphasise this as a major limitation: while high detection accuracy is achieved, understanding why a particular alert was raised is often opaque. Incorporating explainability into hybrid architectures—e.g., latent visualization, feature attribution, decision-path tracing—is an important open research problem, especially for forensic and critical infrastructure scenarios [22, 36, 61].

8.5. Ethical, Privacy, and Regulatory Considerations

Deploying generative–discriminative hybrids raises ethical, privacy, and regulatory concerns. Generative models may inadvertently synthesize personally identifiable information (PII) or replicate sensitive patterns, while discriminative classifiers may perpetuate bias or unfairly flag legitimate behaviours. The intersection of AI, cybersecurity and privacy was discussed in recent surveys. Regulations (such as GDPR, IoT security standards) increasingly require transparency, data minimization, and privacy-preserving operations. Hybrid models must therefore incorporate privacy-preserving techniques (differential privacy, federated learning, encrypted inference) and comply with evolving regulatory frameworks. The governance of synthetic data generation, adverse uses, and ethical risks (e.g., generative models used for social engineering attacks) is also underexplored and poses an urgent research avenue [92].

9. Future Research Directions

9.1. Integrating Foundation Models and Large Multimodal Models

The rise of large foundation models (e.g., LLMs, multimodal transformers) opens new possibilities for cybersecurity applications. Future hybrid architectures may integrate large generative models (e.g., multimodal diffusion for images, logs, binaries) with discriminative decision layers tailored to cyber tasks. Generative modules could leverage pretrained foundation models for synthesis, representation learning, or anomaly detection; discriminative modules would fine-tune these embeddings for classification. Research is needed to adapt large models to constrained cyber environments, ensure robustness, and understand usage patterns. Surveys on security and privacy of large models highlight emerging risks and opportunities [11, 39, 58, 98].

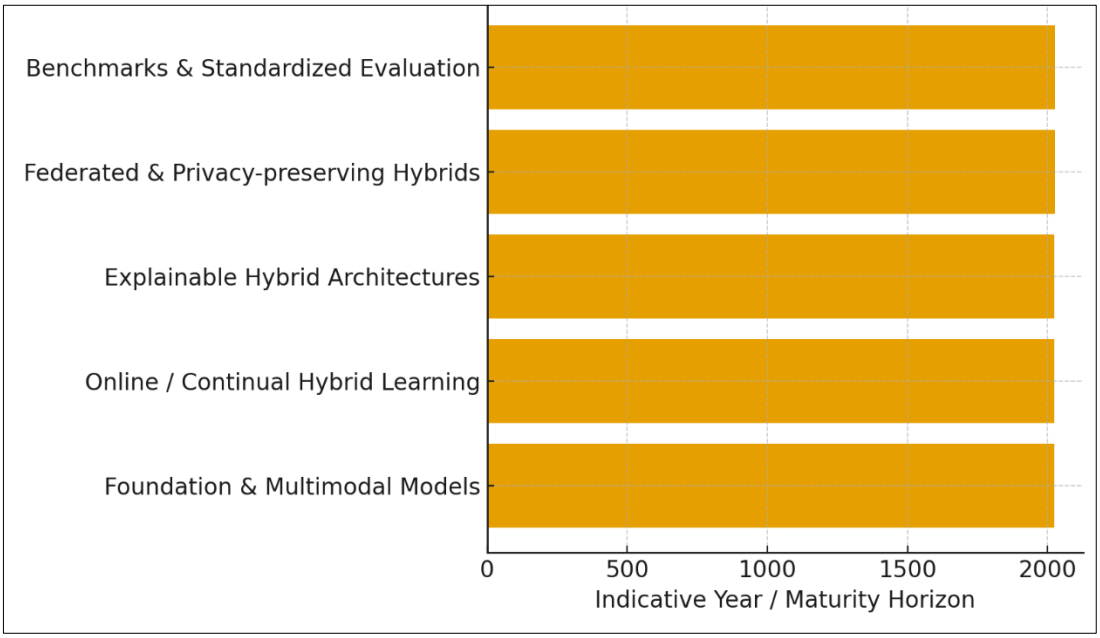

Fig 7: Future Research Agenda for Hybrid Models

9.2. Hybrid Systems with Online and Continual Learning

Real-world cyber threats evolve continuously; static models degrade over time. Future hybrid systems should incorporate online and continual learning, where generative modules adapt to changing data distributions and discriminative modules update incrementally. This may involve self-supervised latent updates, replay buffers, and generative augmentation of new attack patterns as they emerge. Such adaptive hybrids reduce reliance on periodic retraining and improve robustness to concept drift and evolving adversaries [103].

9.3. Explainable Hybrid Architectures

There is a growing need for explainability in cybersecurity ML. Hybrid models present additional complexity due to multi-module design. Future work should focus on designing interpretable hybrid architectures where latent generative representations are transparent, and discriminative decisions are traceable. Approaches could include latent space visualization, attention-based interpretability, attribution across the generative–discriminative boundary, and domain-specific explanation frameworks (e.g., ICS control flows, malware graphs). Surveys on XAI emphasise these directions [68, 91].

9.4. Federated and Privacy-Preserving Hybrid Models

In cloud, IoT and CPS deployments, data locality, privacy and regulatory constraints often prevent centralized model training. Future research should develop hybrid generative–discriminative models that operate in federated or privacy-preserving settings: e.g., federated VAE or GAN modules produce synthetic samples locally, discriminative models aggregate across sites; differential privacy or encrypted inference ensures compliance; generative modules may help anonymize or augment data while preserving utility. Surveys of PPML highlight this gap [33, 72, 96].

9.5. Benchmarks and Standardized Evaluation Protocols

A major bottleneck in current research is the lack of standard benchmarks, evaluation protocols and comparative frameworks for hybrid generative–discriminative cybersecurity models. Future efforts should define standard datasets (with realistic traffic, malware variants, anomaly scenarios), balanced class distributions, adversarial testbeds, drift scenarios, and performance metrics (accuracy, robustness, latency, resource cost, interpretability). Standardization will enable fair comparisons, reproducibility and accelerate adoption in operational environments. The IDS surveys emphasise this need [39, 62, 78].

In summary, the application of hybrid generative–discriminative models spans network IDS, endpoint forensics, IoT/cloud, CPS and threat intelligence domains, offering improved data representation, imbalance handling, adaptability and detection fidelity. However, significant challenges remain including data scarcity, deployment constraints, adversarial robustness, interpretability, and ethical issues. Looking ahead, research that integrates foundation models, supports continual learning, ensures transparency, operates under privacy constraints and uses standardized benchmarks will push the field toward operational readiness and real-world impact.

10. Conclusion

This review demonstrates that hybrid generative–discriminative architectures represent a significant advancement in cybersecurity, offering enhanced detection capabilities, improved generalization to zero-day threats, and superior resilience against adversarial manipulation. By combining the generative capacity to model complex data distributions with the discriminative strength to make accurate and decisive classifications, hybrid models provide a strategically balanced solution for modern cyber defense. Across intrusion detection, malware forensics, IoT/cloud security, fraud analytics, and threat-intelligence applications, these models consistently outperform traditional methods, particularly in handling data imbalance and learning rich, domain-aware representations. However, despite their advantages, challenges remain. Hybrids are computationally demanding, often difficult to interpret, and can themselves be vulnerable to adversarial exploitation. Furthermore, concerns involving privacy, regulatory compliance, and real-time deployment constraints continue to limit large-scale adoption. Addressing these limitations will require sustained

research into explainable architectures, federated and privacy-preserving training mechanisms, scalable deployment pipelines, and unified evaluation benchmarks. Looking forward, integrating hybrid models with foundation-scale multimodal architectures, continual learning mechanisms, and standardized testing frameworks has the potential to reshape cybersecurity into a more adaptive, intelligent, and proactive discipline. This review serves as a foundation for guiding that evolution and informing future system design.

11. References

1. Han X, Zheng H, Zhou M. CARD: Classification and regression diffusion models. In: Advances in Neural Information Processing Systems 35; 2022. p. 18100-18115.
2. Asghari P, Soleimani E, Nazerfard E. Online human activity recognition employing hierarchical hidden Markov models. J Ambient Intell Humaniz Comput. 2020;11(3):1141-1152.
3. Al-Ajlan M, Ykhlef M. A review of generative adversarial networks for intrusion detection systems: Advances, challenges, and future directions. Computers Materials Continua. 2024;81(2):2053-2076. doi: 10.32604/cmc.2024.055891
4. Yinka-Banjo C, Ugot OA. A review of generative adversarial networks and its application in cybersecurity. Artif Intell Rev. 2020;53(3):1721-1736.
5. Halvorsen J, Izurieta C, Cai H, Gebremedhin A. Applying generative machine learning to intrusion detection: A systematic mapping study and review. ACM Comput Surv. 2024;56(10):Article 257. doi: 10.1145/3659575
6. Sajid M, Malik KR, Almogren A, Safdar T, Khan AH, Tanveer J, *et al*. Enhancing intrusion detection: A hybrid machine and deep learning approach. J Cloud Comput. 2024;13:123. doi: 10.1186/s13677-024-00685-x
7. Arnob AKB, Chowdhury RR, Chaiti NA, Saha S, Roy A. A comprehensive systematic review of intrusion detection systems: Emerging techniques, challenges, and future research directions. J Edge Computing. 2025;4(1):73-104. doi: 10.55056/jec.885
8. Feng Y, Si Y, Zhang J, Cai Z, Zhao H. SA-WGAN based data enhancement method for industrial Internet intrusion detection. Computers Materials Continua. 2025;84(3):4431-4449. doi: 10.32604/cmc.2025.064696
9. Ajayi B, Barakat B, McGarry K. Leveraging VAE-derived latent spaces for enhanced malware detection with machine learning classifiers. arXiv preprint arXiv:2503.20803. 2025.
10. Gunduz H. Malware detection framework based on graph variational autoencoder extracted embeddings from API-call graphs. PeerJ Comput Sci. 2022;8:e988.
11. Dong S, *et al*. A-CAVE: Network abnormal traffic detection algorithm based on variational autoencoder. Array. 2023;17:100285. doi: 10.1016/j.array.2023.100285
12. Guo D, Xie Y. Research on network intrusion detection model based on hybrid sampling and deep learning. Sensors. 2025;25(5):1578. doi: 10.3390/s25051578
13. Liu Y, Du H, Niyato D, Kang J, Xiong Z, Kim DI, *et al*. Deep generative model and its applications in efficient wireless network management: A tutorial and case study. IEEE Wirel Commun. 2024;31(4):199-207.
14. Zheng C, Wu G, Bao F, Cao Y, Li C, Zhu J. Revisiting discriminative vs. generative classifiers: Theory and implications. In: International Conference on Machine Learning; 2023 Jul. PMLR; p. 42420-42477.
15. Arifin MM, Ahmed MS, Ghosh TK, Udoy IA, Zhuang J, Yeh JH. A survey on the application of generative adversarial networks in cybersecurity: Prospective, direction and open research scopes. arXiv preprint arXiv:2407.08839. 2024.
16. Dunmore A, Jang-Jaccard J, Sabrina F, Kwak J. A comprehensive survey of generative adversarial networks (GANs) in cybersecurity intrusion detection. IEEE Access. 2023;11:84241-84270. doi: 10.1109/ACCESS.2023.3296707
17. Mahmoudi I, *et al*. Toward generative AI-based intrusion detection systems for Internet of Vehicles. Future Internet. 2025;17(7):310. doi: 10.3390/fi17070310
18. Halvorsen J, Izurieta C, Cai H, Gebremedhin A. Applying generative machine learning to intrusion detection: A systematic mapping study and review. ACM Comput Surv. 2024;56(10):Article 257. doi: 10.1145/3659575
19. Ravanmehr R, Mohamadrezaei R. Deep learning overview. In: Session-Based Recommender Systems Using Deep Learning. Cham: Springer Nature Switzerland; 2023. p. 27-72.
20. Hozouri A, *et al*. A comprehensive survey on intrusion detection systems: Emerging techniques, challenges, and future research directions. Sensors. 2025;25(20):6261. doi: 10.3390/s25206261
21. Argouarc'h E, Desbouvries F, Barat E, Kawasaki E. Generative vs. discriminative modeling under the lens of uncertainty quantification. arXiv preprint arXiv:2406.09172. 2024.
22. Saranya S, *et al*. Generative AI-based intrusion detection of imbalanced cyber device networks using GAN. Int J Comput Appl. 2025;187(15):1-9.
23. Siqueira LP, *et al*. A comprehensive survey on intrusion detection systems in the era of AI. Sensors. 2025;25(20):6261.
24. Makhlouf IB, Kilani G, Jaafar F, Nakouri H. A hybrid approach to improve the intrusion detection systems using generative artificial intelligence and deep reinforcement learning. 2025.
25. Ajayi B, Barakat B, McGarry K. Leveraging VAE-derived latent spaces for enhanced malware detection with machine learning classifiers. arXiv preprint arXiv:2503.20803. 2025.
26. Ambekar NG, Thokchom S. UL-VAE: An unsupervised learning approach for zero-day malware detection using variational autoencoder. In: 2024 International Conference on Computational Intelligence and Network Systems (CINS); 2024. p. 1-7. doi: 10.1109/CINS63881.2024.10864450
27. Nalisnick E, Matsukawa A, Teh YW, Lakshminarayanan B. Hybrid models with deep and invertible features. In: Proceedings of the 36th International Conference on Machine Learning; 2019. p. 4723-4732.
28. Wang Y, Zhang B, Zhang Y, Li H. Industrial Internet intrusion detection method based on VAE-WGAN-GP data enhancement. Bull Cherkasy State Technol Univ. 2024;30(1):41-50.
29. Tian W, Shen Y, Guo N, Yuan J, Yang Y. VAE-

WACGAN: An improved data augmentation method based on VAEGAN for intrusion detection. Sensors. 2024;24(18):6035. doi: 10.3390/s24186035
30. Gul S, Arshad J, *et al*. WGAN-DL-IDS: An efficient framework for intrusion detection system using WGAN, random forest, and deep learning approaches. Computers. 2024;14(1):4.
31. Halvorsen J, Izurieta C, Cai H, Gebremedhin A. Applying generative machine learning to intrusion detection: A systematic mapping study and review. ACM Comput Surv. 2024;56(10):Article 257. doi: 10.1145/3659575
32. Alharbi A, Alsulaiman F, El Saddik A. Inpainting forgery detection using hybrid generative-discriminative models. Appl Comput Inform. 2024;20(1-2):89-103. doi: 10.1108/ACI-03-2023-0063
33. Kang F, Zhang H, Zhang Y. M2M-VAEGAN-IDS: VAE-GAN-guided cross-class generation for imbalanced intrusion detection datasets. Electronics. 2025;14(11):2103. doi: 10.3390/electronics14112103
34. Deldar A, Abadi M. Deep learning for zero-day malware detection and classification: A survey. Comput Netw. 2024;243:110476. doi: 10.1016/j.comnet.2024.110476
35. Wasswa H, Li W, Li Q. Latent space alignment for robust detection of IoT botnet attacks. Knowl-Based Syst. 2025;307:112865. doi: 10.1016/j.knosys.2025.112865
36. Hernadi V. Improving network intrusion detection of minority classes using GAN-based data augmentation [Master's thesis]. Linnaeus University; 2024.
37. Sehsah AI, Darwish A, *et al*. A hybrid variational autoencoder and Wasserstein GAN with gradient penalty for anomaly detection in distance matrices. Sci Rep. 2025;15:12345. doi: 10.1038/s41598-025-94747-y
38. Riaz R, Hussain F, *et al*. A novel ensemble Wasserstein GAN framework for minority-class intrusion detection. Sci Rep. 2025;15:7654. doi: 10.1038/s41598-025-07533-1
39. Sarhan M, Layeghy S, Portmann M. From zero-shot machine learning to zero-day attack detection. J Netw Comput Appl. 2023;216:103706. doi: 10.1016/j.jnca.2023.103706
40. Kamal H, Mashaly M. Robust intrusion detection system using an improved hybrid deep learning model with explainability. Technologies. 2025;13(3):102. doi: 10.3390/technologies13030102
41. Ambekar NG, Thokchom S. UL-VAE: An unsupervised learning approach for zero-day malware detection using variational autoencoder. In: 2024 International Conference on Computational Intelligence and Network Systems (CINS). IEEE; 2024.
42. Taylor T. Using variational autoencoders with machine learning for intrusion detection. Artif Intell Adv. 2025;3(1):1-15.
43. Li Z, *et al*. An intrusion detection method combining variational auto-encoder and generative adversarial networks. Comput Netw. 2024;242:110724. doi: 10.1016/j.comnet.2024.110724
44. Duc MV, *et al*. Detecting emerging DGA malware in federated environments via variational autoencoder-based clustering and resource-aware client selection. Future Internet. 2025;17(7):299. doi: 10.3390/fi17070299
45. Choi A, Trousil R, *et al*. Synthetic malware using deep variational autoencoders and generative adversarial networks. EAI Endorsed Trans Internet Things. 2024;10(3):e5. doi: 10.4108/eetiot.6566
46. Arifin MM, Ahmed MS, Ghosh TK, Udoy IA, Zhuang J, Yeh JH. A survey on the application of generative adversarial networks in cybersecurity: Prospective, direction and open research scopes. arXiv preprint arXiv:2407.08839. 2024.
47. 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. doi: 10.1109/ACCESS.2023.3296707
48. Al-Ajlan M, Ykhlef M. A review of generative adversarial networks for intrusion detection systems: Advances, challenges, and future directions. Computers Materials Continua. 2024;81(2):2053-2076. doi: 10.32604/cmc.2024.055891
49. Yang Y, *et al*. A CE-GAN based approach to address data imbalance in intrusion detection systems. Sci Rep. 2025;15:90815. doi: 10.1038/s41598-025-90815-5
50. Gebrehans G, *et al*. Generative adversarial networks for dynamic malware detection. IEEE Trans Artif Intell. 2025;6(8):1234-1248.
51. Redhu A, *et al*. Deep learning-powered malware detection in cyberspace: A review. Front Phys. 2024;12:1349463.
52. Tang B, Lu Y, Li Q, Bai Y, Yu J, Yu X. A diffusion model based on network intrusion detection method for industrial cyber-physical systems. Sensors. 2023;23(3):1141. doi: 10.3390/s23031141
53. Cai S, Zhao Y, Lyu J, Wang S, Hu Y, Cheng M, *et al*. DDP-DAR: Network intrusion detection based on denoising diffusion probabilistic model and dual-attention residual network. Neural Netw. 2025;184:107064.
54. Merzouk MA, Beurier E, Yaich R, Boulahia-Cuppens N, Cuppens F, Khomh F. Diffusion-based adversarial purification for intrusion detection. In: IFIP Annual Conference on Data and Applications Security and Privacy; 2025 Jun. Cham: Springer Nature Switzerland; p. 351-370.
55. Bao K, Trousil M. Leveraging generative artificial intelligence for intrusion detection systems. GSC Adv Res Rev. 2024;18(3):217-230.
56. Ocheredko AR, *et al*. A method for determining the credibility of intrusion detection system alerts using Bayesian models. In: Proceedings of the 20th International Conference on Security and Cryptography; 2023. p. 457-468.
57. Kim Y, Dán G, Zhu Q. Human-in-the-loop cyber intrusion detection using active learning. IEEE Trans Inf Forensics Secur. 2024.
58. Carrodano C. Data-driven risk analysis of nonlinear factor interactions in road safety using Bayesian networks. Sci Rep. 2024;14(1):18948.
59. Liu L, *et al*. A network intrusion detection method based on contrastive learning and Bayesian Gaussian mixture model. Cybersecurity. 2025;8(1):15. doi: 10.1186/s42400-025-00364-7
60. Wu K, *et al*. A network intrusion detection method incorporating Bayesian attack graph and XGBoost incremental learning. Future Internet. 2023;15(4):128. doi: 10.3390/fi15040128

61. Li Y, *et al*. A comprehensive survey on intrusion detection algorithms: Classical, machine learning and deep learning approaches. Comput Electr Eng. 2025;118:109863. doi: 10.1016/j.compeleceng.2024.109863
62. Hartono DD. Efficient GAN-based adversarial example generation for intrusion detection systems [Master's thesis]. California Polytechnic State University; 2025.
63. Zhao S, *et al*. Adversarial attack against black-box IDS using generative adversarial networks. Procedia Comput Sci. 2021;187:67-74. doi: 10.1016/j.procs.2021.04.037
64. Lin W, *et al*. IDSGAN: Generative adversarial networks for attack generation against intrusion detection. In: Cyber Security. Springer; 2022. p. 121-139.
65. Hartono DD. Efficient Gan-Based Adversarial Example Generation Against Ml-Based Network Intrusion Detection Systems. 2025.
66. Hozouri A, Mirzaei A, Effatparvar M. A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges. Discov Artif Intell. 2025;5(1):314.
67. Zhang Y, Xu Y, Li Z, Wang L. A review of deep learning applications in intrusion detection systems. Appl Sci. 2025;15(3):1552. doi: 10.3390/app15031552
68. Rahman MM, *et al*. A survey on intrusion detection system in IoT networks. Internet Things Cyber-Phys Syst. 2025;5:100-122. doi: 10.1016/j.iotcps.2024.100122
69. Kaissar A, Bou Nassif A, Injadat M. A survey on network intrusion detection using convolutional neural network. ITM Web Conf. 2022;43:01003. doi: 10.1051/itmconf/20224301003
70. AbdelHalim AP. Deep learning techniques for network intrusion detection: A comprehensive study. Int J Intell Comput Inf Sci. 2025;25(2):45-60.
71. Liu D, Zheng X, Wang P, Jiao W. Deep learning-based intrusion detection: A CNN-LSTM-Transformer approach for enhanced network security. In: Proceedings of the 10th International Conference on Cyber Security and Information Engineering (ICCSIE 2025). doi: 10.1145/3759179.3760451
72. Afraji DMAA, Lloret J, Peñalver L. An integrated hybrid deep learning framework for intrusion detection in IoT and IIoT networks using CNN-LSTM-GRU architecture. Computation. 2025;13(9):222.
73. Doost PA, Moghadam A, *et al*. A new intrusion detection method using ensemble classification and feature selection. Sci Rep. 2025;15:Article 98604. doi: 10.1038/s41598-025-98604-w
74. Ghadami R, *et al*. An intrusion detection system in the Internet of Things with GAN-based balancing and parallel CNN-LSTM detection. Sci Rep. 2025;15:Article 22074-3. doi: 10.1038/s41598-025-22074-3
75. Farhan BI, *et al*. Survey of intrusion detection using deep learning in the Internet of Things. J Netw Comput Appl. 2021;190:103139.
76. Kanimozhi R, *et al*. Deep reinforcement learning-based intrusion detection scheme for software-defined networking. Sci Rep. 2025;15:24869. doi: 10.1038/s41598-025-24869-w
77. Abdulameer HS. IoT intrusion detection using transformer-based anomaly learning. J Al-Qadisiyah Comput Sci Math. 2025;17(3):278-291. doi: 10.29304/jqcsm.2025.17.3.2432
78. Almadhor A, *et al*. Evaluating large transformer models for anomaly detection in IoT networks. Sci Rep. 2025;15:21826-5. doi: 10.1038/s41598-025-21826-5
79. Alshomrani M, *et al*. Survey of transformer-based malicious software detection techniques. Electronics. 2024;13(23):4677. doi: 10.3390/electronics13234677
80. Ashraf MWA. Detecting cyber threats in Internet of Things using a hybrid SVM rule-based detection method (HSVMR-D) [Ph.D. thesis]. University of Hagen; 2024.
81. Abdulhameed AA, *et al*. An optimized model for network intrusion detection in the cloud environment. Mesopotamian J CyberSecurity. 2024;3(1):1-15.
82. Alsolami T, Alsharif B, Ilyas M. Enhancing cybersecurity in healthcare: Evaluating ensemble learning models for intrusion detection in the internet of medical things. Sensors. 2024;24(18):5937.
83. Sanmorino A. Improving threat detection in information security with ensemble learning. ACIG J Technol Secur. 2025;1(2):45-60.
84. Bibers I, *et al*. An ensemble learning framework for enhancing AI-based intrusion detection systems. Appl Sci. 2025;15(19):10579. doi: 10.3390/app151910579
85. Dhanvijay DM, *et al*. Cyber intrusion detection using ensemble of deep learning and traditional models in IoT. Internet Things Cyber-Phys Syst. 2025;6:1-15. doi: 10.1016/j.iotcps.2025.100145
86. Zhong M, Lin M, Zhang C. A survey on graph neural networks for intrusion detection systems: Methods, trends and challenges. Comput Secur. 2024;141:103821. doi: 10.1016/j.cose.2024.103821
87. Wang Y, *et al*. BS-GAT: A network intrusion detection system based on graph neural network for edge computing. Cybersecurity. 2025;8(1):5. doi: 10.1186/s42400-024-00296-8
88. Jahin MA, *et al*. Graph neural networks for network intrusion detection: Architectures and data augmentation. arXiv preprint arXiv:2503.00961. 2025.
89. Zhang A, *et al*. ResACAG: A graph neural network-based intrusion detection system with residual adaptive context-aware graph. Comput Electr Eng. 2025;121:109513. doi: 10.1016/j.compeleceng.2024.109513
90. Galli D, Venturi A, Marasco I, Marchetti M. Evaluating explainability of graph neural networks for network intrusion detection with structural attacks. In: Proceedings of ARES 2025 Workshops (CEUR-WS Vol. 3962).
91. Latif Martínez H, *et al*. Graph attention networks for contextual anomaly detection in network traffic (GAT-AD) [Master's thesis]. Universitat Politècnica de Catalunya; 2025.
92. Li Z, Wu X, Liu H, Liu Z. An intrusion detection method combining variational auto-encoder and generative adversarial networks. Comput Netw. 2024;242:110724. doi: 10.1016/j.comnet.2024.110724
93. Wang Y, Zhang Z. Industrial Internet intrusion detection method based on VAE-WGAN-GP data enhancement. Acad J Comput Inf Sci. 2024;7(1):36-44.
94. Kang F, Zhang H, Zhang Y. VAE-GAN-guided cross-class generation: A class imbalance data augmentation method for network intrusion detection (M2M-VAEGAN-IDS). Electronics. 2025;14(11):2103. doi: 10.3390/electronics14112103

95. Rahman MA, *et al*. Leveraging GANs for synthetic data generation to improve intrusion detection. FAITH Future AI Technol Harmony. 2025;1(1):1-18.
96. Klinkhamhom C, *et al*. MIDS-GAN: Minority intrusion data synthesizer using generative adversarial networks. Mathematics. 2025;13(21):3391. doi: 10.3390/math13213391
97. Ambekar NG, Thokchom S. UL-VAE: An unsupervised learning approach for zero-day malware detection using variational autoencoder. In: 2024 International Conference on Computational Intelligence and Network Systems (CINS); 2024. p. 1-7. doi: 10.1109/CINS63881.2024.10864450
98. Ajayi B, Barakat B, McGarry K. Leveraging VAE-derived latent spaces for enhanced malware detection with machine learning classifiers. arXiv preprint arXiv:2503.20803. 2025.
99. Kim JY, Shin Y, Mohaisen A, Kim HK. Obfuscated malware detection using deep generative models. Comput Secur. 2022;112:102507. doi: 10.1016/j.cose.2021.102507
100. Gunduz H, Das R. Malware detection framework based on graph variational autoencoder and deep neural network. Secur Commun Netw. 2022;2022:1-11. doi: 10.1155/2022/7811021
101. Bekkaye C, *et al*. Generative hybrid models for fraud detection in auto insurance with a comparative analysis of VAE, GAN, and diffusion approaches. Discov Artif Intell. 2025;5:574. doi: 10.1007/s44163-025-00574-5
102. Strelcenia E, Prakoonwit S. A survey on GAN techniques for data augmentation to address the imbalanced data issues in credit card fraud detection. Mach Learn Knowl Extr. 2023;5(2):304-329. doi: 10.3390/make5020016
103. RezvaniNejad M, *et al*. WDAE-GAN: A hybrid dual autoencoder and generative adversarial network framework for anomaly detection. Expert Syst Appl. 2025;256:125612. doi: 10.1016/j.eswa.2025.125612
104. Shirazi H, *et al*. Improved phishing detection algorithms using adversarial autoencoder synthesized data. In: 2021 IEEE International Conference on Big Data (Big Data); 2021. p. 649-656.
105. Elberri MA, *et al*. A cyber defense system against phishing attacks with deep learning and meta-heuristic optimization. Int J Inf Secur. 2024. doi: 10.1007/s10207-024-00851-x
106. Jabir R, *et al*. Phishing attacks in the age of generative artificial intelligence. AI. 2025;6(8):174. doi: 10.3390/ai6080174
107. Lin Z, Shi Y, Xue Z. IDSGAN: Generative adversarial networks for attack generation against intrusion detection. In: Information and Communications Security. Springer; 2018. p. 18-32. doi: 10.1007/978-3-030-05981-0_7
108. Zhao S, *et al*. Adversarial attack against black-box IDS using generative adversarial networks. Procedia Comput Sci. 2021;187:67-74. doi: 10.1016/j.procs.2021.04.037
109. Ndayipfukamiye B, *et al*. A systematic review of GANs for threat detection and response. arXiv preprint arXiv:2509.20411. 2025.
110. Ahmed M, *et al*. Network intrusion detection systems: Machine learning, deep learning, and adversarial robustness. Al-Salam J Eng Technol. 2025;5(1):1-27.