10.48047/jocaaa.2022.30.02.76

Adversarial Machine Learning in Intrusion Detection Systems: Threats and Mitigations

Nitin Bodade
nitinbodade@gmail.com

Abstract*—Network Intrusion Detection Systems (IDS) built on machine learning and deep learning models have become the backbone of modern cyber defense, offering the ability to detect novel and previously unseen attacks that signature-based systems cannot recognize. However, the same statistical learning properties that give these models their detection power also make them susceptible to adversarial machine learning (AML), a class of techniques that deliberately crafts perturbed inputs to cause misclassification while preserving the functional validity of the underlying network traffic. This paper presents a comprehensive examination of the adversarial threat landscape confronting learning-based IDS, spanning evasion attacks at inference time, data poisoning attacks during training, model extraction and inference attacks against deployed classifiers, and the unique problem-space constraints that separate network security from the image-domain origins of adversarial machine learning research. We develop a structured taxonomy of attacks according to attacker knowledge, attacker goal, and attack timing, and we review the principal attack methodologies reported in the literature, including gradient-based feature perturbation, generative adversarial network based traffic synthesis, and black-box query attacks. We then examine the corresponding defense landscape, covering adversarial training, defensive distillation, ensemble and moving-target defenses, feature squeezing, and certified robustness approaches, while highlighting the persistent gap between reported robustness and real-world resilience. The paper concludes by identifying open research challenges, including the absence of standardized adversarial robustness benchmarks for network intrusion detection, the tension between real-time detection constraints and computationally expensive defenses, and the need for problem-space-aware evaluation methodologies that respect the semantic and protocol constraints of real network traffic.*

Keywords*—Adversarial Machine Learning, Intrusion Detection Systems, Evasion Attacks, Data Poisoning, Adversarial Training, Network Security, Deep Learning Robustness, Generative Adversarial Networks*

I. INTRODUCTION

Cybersecurity threats have grown in both volume and sophistication over the past decade, driven by the expansion of cloud computing, the proliferation of Internet of Things (IoT) devices, and the increasing monetary incentives behind cybercrime. Traditional signature-based intrusion detection systems, which rely on predefined rules and known attack fingerprints, are increasingly unable to keep pace with the volume of zero-day exploits and polymorphic malware that characterize contemporary threat campaigns [1]. In response, the security community has turned to machine learning (ML) and deep learning (DL) techniques, which learn statistical patterns of normal and malicious behavior directly from network traffic data rather than relying on hand-crafted rules [2]. Architectures such as convolutional neural networks, recurrent neural networks, autoencoders, and ensemble classifiers have been shown to achieve high detection accuracy on benchmark datasets such as NSL-KDD, CICIDS2017, and UNSW-NB15, often exceeding ninety-five percent accuracy under standard evaluation protocols [3].

Despite these encouraging results, the same properties that allow machine learning models to generalize from data also expose them to a distinct category of vulnerability. Seminal work in computer vision demonstrated that deep neural networks can be reliably fooled by small,

10.48047/jocaaa.2022.30.02.76

carefully constructed perturbations added to an input image, perturbations that are often imperceptible to a human observer yet cause a classifier to output an incorrect label with high confidence [4]. Goodfellow et al. formalized this phenomenon through the Fast Gradient Sign Method (FGSM), showing that the apparent brittleness of neural networks arises from their approximately linear behavior in high-dimensional input spaces rather than from any inherent weakness of a specific architecture [5]. This body of research, now broadly referred to as adversarial machine learning, has since expanded well beyond image classification into speech recognition, natural language processing, and, increasingly, network security.

The transfer of adversarial machine learning concepts into the intrusion detection domain raises concerns that are qualitatively different from those in computer vision. An intrusion detection system does not operate in a benign or neutral environment; it is deployed precisely because an adaptive, motivated adversary is attempting to evade it. Any attacker capable of observing or approximating the decision behavior of a deployed IDS has a direct incentive to search for adversarial perturbations that convert malicious traffic into a form that is classified as benign, thereby defeating the very purpose of the defense [6]. Unlike pixel-level perturbations in an image, however, adversarial perturbations to network traffic must respect protocol semantics, feature dependencies, and functional constraints; an attacker cannot arbitrarily alter a packet's header fields or payload without breaking the underlying communication or invalidating the intended malicious functionality. This tension between statistical evasion and functional validity, often referred to as the problem-space versus feature-space distinction, is central to understanding why adversarial attacks against IDS require specialized treatment rather than a direct import of computer-vision techniques.

This paper makes four principal contributions. First, it develops a structured taxonomy of adversarial attacks against learning-based intrusion detection systems, organized along the dimensions of attacker knowledge, attacker objective, and attack timing. Second, it surveys the principal attack methodologies reported in the literature, including gradient-based perturbation methods, generative adversarial network based traffic synthesis, and black-box query-based attacks, along with representative quantitative findings regarding their effectiveness against deployed classifiers. Third, it reviews the corresponding landscape of defense and mitigation strategies, evaluating their strengths, limitations, and susceptibility to adaptive counter-attacks. Fourth, it identifies open research challenges that remain insufficiently addressed, including the lack of standardized benchmarks for adversarial robustness evaluation in the network security domain and the practical tension between defense overhead and the real-time performance requirements of production IDS deployments. The remainder of this paper is organized as follows: Section II reviews related work; Section III introduces the fundamental concepts of adversarial machine learning; Section IV presents the threat model and attack taxonomy; Section V surveys attack methodologies specific to IDS; Section VI discusses defense and mitigation strategies; Section VII presents an experimental evaluation framework and discussion of reported results; Section VIII outlines open challenges and future directions; and Section IX concludes the paper.

II. LITERATURE REVIEW

The study of adversarial vulnerabilities in machine learning predates its recent popularity in deep learning research. Biggio and Roli trace the origins of adversarial machine learning to earlier work on evading linear classifiers used in spam filtering and malware detection, arguing that security-sensitive applications of machine learning have always implicitly operated under an adversarial threat model even before the term 'adversarial examples' was coined in the deep learning literature [6]. Their retrospective analysis, often cited as one of the foundational surveys of the field, distinguishes between evasion attacks that occur at test time and poisoning

10.48047/jocaaa.2022.30.02.76

attacks that corrupt the training process, a distinction that remains central to the taxonomy adopted throughout this paper.

Within the specific domain of network intrusion detection, Rigaki was among the first to systematically demonstrate that adversarial perturbation techniques originally developed for image classifiers could be adapted to degrade the performance of machine-learning-based network intrusion detectors, showing that classifiers trained on NSL-KDD features experienced substantial accuracy degradation when exposed to perturbed feature vectors [7]. Building on this line of work, Yang et al. applied both the Jacobian-based Saliency Map Attack and FGSM to deep learning IDS models, reporting that carefully selected feature perturbations could reduce detection rates significantly while preserving a subset of features that are difficult for an attacker to modify without breaking traffic validity [8]. Usama et al. extended this approach using generative adversarial networks, proposing a framework in which a generator network learns to synthesize adversarial network flow features that a discriminator, standing in for the target IDS, is unable to distinguish from benign traffic, achieving high evasion rates against several classifier architectures [9].

Apruzzese et al. conducted a broader empirical evaluation of deep learning based IDS robustness, systematically testing several classifier architectures against a range of perturbation strategies and concluding that even models with strong reported accuracy on clean test data suffered dramatic performance collapse under adversarial conditions, and that robustness varied considerably depending on which subset of features the attacker was permitted to modify [10]. Alhajjar et al. proposed a GAN-based black-box evasion framework capable of generating adversarial network traffic without requiring access to the internal parameters of the target IDS, demonstrating that black-box attacks, while generally less effective than white-box attacks, remained a practical and realistic threat against commercial-grade detection systems [11].

Sheatsley et al. addressed the problem-space constraint issue directly, introducing a method for generating adversarial examples that explicitly respects feature dependencies and domain constraints inherent to network traffic, an important departure from earlier work that largely ignored the feasibility of the generated perturbations in real deployed networks [12]. Ibitoye et al. provided a threat-modeling perspective specific to IoT network intrusion detection, cataloging the practical attack surfaces available to an adversary targeting resource-constrained IoT gateways and highlighting that lightweight IDS models deployed on such devices are often more vulnerable due to simplified feature sets [13]. Anthi et al. extended adversarial evaluation to IoT-specific datasets, confirming that adversarial vulnerability is not confined to conventional enterprise network IDS but extends broadly across deployment contexts [14].

Debicha et al. examined the transferability of adversarial examples across different IDS architectures, an important practical consideration because an attacker rarely has complete knowledge of the exact model deployed by a defender; their results indicated that adversarial examples crafted against one architecture frequently transfer with reduced but non-trivial effectiveness to structurally different classifiers, reinforcing the black-box threat model as realistic [15]. Han et al. proposed a moving target defense strategy that periodically alters the feature space or model ensemble used by the IDS specifically to reduce the shelf life of any adversarial example an attacker might have crafted, representing one of the few defense-oriented contributions that explicitly assumes an adaptive, persistent adversary [16]. Collectively, this body of literature establishes that adversarial vulnerability in learning-based IDS is well documented empirically, yet the surveyed works also reveal recurring gaps: a majority of studies still adapt image-domain attack algorithms with only partial attention to problem-space feasibility, defense evaluations are frequently conducted against a narrow set of known attacks rather than adaptive adversaries, and standardized benchmarks comparable to

10.48047/jocaaa.2022.30.02.76

those available in the computer vision robustness literature remain largely absent from the network security domain. These gaps motivate the structured treatment developed in the remainder of this paper.

III. BACKGROUND: FUNDAMENTALS OF ADVERSARIAL MACHINE LEARNING

Before examining attacks specific to intrusion detection, it is useful to establish the fundamental vocabulary of adversarial machine learning. An adversarial example is an input that has been deliberately modified, typically through the addition of a small perturbation, so that a target model produces an incorrect output while the modified input remains close, under some distance metric, to a legitimate original input [4]. The perturbation is commonly bounded by an epsilon value defining an allowable region, often described geometrically as an epsilon-ball around the original input in feature space, within which the attacker is permitted to search for a perturbation that changes the model's prediction. The choice of distance metric, typically an L0, L2, or L-infinity norm in image-domain research, does not translate directly to tabular or flow-based network data, where features may be categorical, bounded by protocol semantics, or mutually dependent, a distinction elaborated further below.

Adversarial threat models are conventionally described along three orthogonal dimensions. The first dimension concerns attacker knowledge: a white-box attacker has complete access to the target model's architecture, parameters, and training data, allowing exact computation of gradients with respect to the loss function; a black-box attacker has no such access and must rely on querying the model and observing its outputs, or on transferability from a locally trained substitute model; and a gray-box attacker occupies an intermediate position, possessing partial knowledge such as the feature set or general architecture family without exact parameter access [6]. The second dimension concerns attacker objective: an untargeted attack seeks merely to cause any misclassification, such as converting a 'malicious' label to any label other than malicious, whereas a targeted attack seeks a specific incorrect label, such as forcing a particular attack category to be classified specifically as 'benign' traffic. The third dimension concerns attack timing: evasion attacks manipulate inputs at inference time against an already-trained model, while poisoning attacks manipulate the training data or the online learning process itself, degrading the model before it is ever deployed or corrupting it incrementally during continuous retraining.

Several algorithmic techniques form the foundation of modern adversarial example generation. The Fast Gradient Sign Method computes the gradient of the model's loss function with respect to the input and perturbs the input in the direction that maximizes the loss, subject to an L-infinity budget, producing adversarial examples in a single computationally efficient step [5]. The Projected Gradient Descent (PGD) method extends this idea into an iterative multi-step optimization procedure, repeatedly taking small gradient steps and projecting the result back into the allowed perturbation region, and is widely regarded in the computer vision literature as one of the strongest first-order attacks against which robustness should be evaluated [17]. The Carlini-Wagner attack formulates adversarial example generation as an explicit optimization problem that minimizes perturbation magnitude subject to a misclassification constraint, and has been shown to defeat several defenses that appeared robust against weaker attacks such as FGSM [18]. The Jacobian-based Saliency Map Attack instead computes a saliency map identifying which input features have the greatest influence on the target model's output and perturbs only a minimal subset of the most influential features, a property that is particularly relevant to network intrusion detection because it naturally favors sparse perturbations that are easier to reconcile with real-world traffic constraints [8].

10.48047/jocaaa.2022.30.02.76

The critical distinction that separates adversarial machine learning in network security from its origins in computer vision is the problem-space versus feature-space constraint. In an image classification task, an attacker who perturbs pixel values in feature space directly produces a corresponding valid image in the problem space; the transformation between the two spaces is essentially the identity function. In network intrusion detection, by contrast, the features consumed by a classifier, such as flow duration, byte counts, protocol flags, or inter-arrival time statistics, are derived from underlying raw network traffic through a feature extraction pipeline, and an attacker cannot simply set arbitrary values for these derived features without first determining whether any corresponding valid network traffic could produce them [19]. Pierazzi et al. formalize this requirement, arguing that a realistic problem-space attack must satisfy four properties: the transformation must be available to the attacker, it must preserve the semantics and functionality of the original malicious activity, it must be robust to the specific preprocessing pipeline used by the defender, and the resulting modification must plausibly occur in real network conditions [19]. Attacks that ignore these constraints, however successful in feature space, may correspond to no achievable real-world traffic and therefore overstate the practical threat, a methodological concern that recurs throughout the literature surveyed in this paper.

IV. THREAT MODEL AND TAXONOMY OF ADVERSARIAL ATTACKS ON IDS

Building on the foundational concepts introduced above, adversarial attacks against intrusion detection systems can be organized into a structured taxonomy spanning five principal categories, distinguished by the stage of the machine learning pipeline they target and the objective pursued by the attacker.

Evasion attacks represent the most extensively studied category and occur when an attacker modifies malicious network traffic, or the feature representation derived from it, so that a deployed classifier assigns it a benign label at inference time. The attacker's goal in this category is purely to avoid detection for a specific malicious session, without altering the underlying model itself. Evasion attacks are further subdivided according to attacker knowledge: white-box evasion assumes the attacker has full access to model gradients, an assumption realistic for insider threats or scenarios where the defender uses a publicly documented open-source IDS architecture; black-box evasion assumes only query access or the ability to observe detection outcomes, a more realistic assumption for external attackers targeting commercial or proprietary systems; and gray-box evasion assumes partial knowledge, such as awareness of the feature extraction scheme without knowledge of model parameters, which is a common and pragmatic middle ground adopted in much of the empirical literature [7], [11].

Poisoning attacks target the training phase rather than inference, and involve injecting maliciously crafted samples into the data used to train or retrain an IDS model. Because many production intrusion detection systems periodically retrain on newly observed traffic to adapt to evolving threats, an attacker with the ability to inject traffic into the network, even traffic that does not itself constitute an attack, can gradually shift the decision boundary of the classifier to misclassify a targeted class of future attacks as benign. Poisoning attacks are typically categorized as either availability attacks, which seek to degrade overall model accuracy indiscriminately, or targeted attacks, which seek to induce a specific, narrow misclassification while leaving overall accuracy largely intact so that the compromise remains undetected through aggregate performance monitoring [6].

Model extraction, or model stealing, attacks aim to reconstruct a functionally equivalent substitute of a deployed IDS through repeated querying, without direct access to its parameters.

10.48047/jocaaa.2022.30.02.76

An attacker who successfully extracts a substitute model gains the ability to conduct white-box attacks offline against the substitute, then transfer the resulting adversarial examples to the original target, exploiting the transferability property documented extensively in the literature [15]. Model extraction is of particular concern for IDS vendors who expose their detection logic as a queryable service or embed it in distributable software, since the extraction process effectively converts a black-box deployment into a de facto white-box one from the attacker's perspective.

Membership inference and related privacy attacks constitute a less-studied but increasingly relevant category, in which an attacker seeks to determine whether a specific traffic sample, potentially corresponding to sensitive organizational data, was part of the training set used to build the IDS, or to reconstruct approximate statistical properties of that training data. While less directly damaging to detection efficacy than evasion or poisoning, these attacks raise data governance and compliance concerns for organizations that train IDS models on proprietary or regulated network traffic.

Finally, exploratory or reconnaissance attacks involve an adversary systematically probing a deployed IDS, for example by sending traffic samples near known decision boundaries and observing whether they are flagged, in order to map the approximate shape of the decision surface without necessarily executing an actual evasion or poisoning attack in the same session. Such reconnaissance often serves as a precursor stage that informs a subsequent, more targeted evasion or poisoning campaign, and detecting this probing behavior itself has been proposed as an early-warning defensive signal [13]. Across all five categories, the practical severity of a given attack depends heavily on whether the attacker's manipulation remains realizable in the problem space, underscoring the recurring importance of the constraint discussed in Section III.

V. ATTACK METHODOLOGIES AGAINST INTRUSION DETECTION SYSTEMS

Empirical studies evaluating adversarial attacks against IDS classifiers have converged on several representative methodologies, each illustrating a different point along the trade-off between attack strength, computational cost, and the attacker's assumed level of access. Gradient-based feature perturbation methods, adapting FGSM, PGD, and JSMA directly from the image domain, remain the most common approach in white-box evaluation studies. Applied to classifiers trained on NSL-KDD, CICIDS2017, or UNSW-NB15 feature sets, these methods compute perturbations with respect to a subset of features deemed modifiable by the attacker, typically excluding fields such as protocol type or fixed identifiers that cannot be altered without breaking the underlying connection, while permitting modification of statistical features such as packet counts, byte totals, or flow duration that an attacker can influence through pacing or padding strategies [8], [20].

Reported results across this literature consistently show substantial degradation in detection performance under white-box gradient attacks. Several studies report that convolutional and recurrent neural network based IDS classifiers, which achieve detection accuracies in the high nineties on clean held-out test data, experience accuracy drops to below thirty percent, and in some configurations below ten percent, when evaluated against adversarially perturbed traffic constrained to modifiable features only [10], [21]. This magnitude of degradation illustrates that reported clean-data accuracy figures, which dominate much of the applied IDS literature, provide limited assurance regarding real-world resilience against a motivated adversary, and argues for adversarial robustness to be reported as a standard evaluation metric alongside accuracy, precision, and recall.

10.48047/jocaaa.2022.30.02.76

Generative adversarial network based approaches represent a second major methodology, exemplified by frameworks such as IDSGAN, in which a generator network is trained to transform malicious traffic feature vectors into adversarial variants that a discriminator network, trained to emulate the target IDS's decision function, classifies as benign [9]. This approach has the practical advantage of producing a reusable generator capable of synthesizing large volumes of evasive traffic variants after a single training procedure, rather than requiring per-sample optimization as in gradient-based methods, and several studies report that GAN-generated adversarial traffic achieves evasion success rates exceeding eighty percent against a range of target classifier architectures, including models the generator was not directly trained against, indicating meaningful cross-architecture transferability [9], [11].

Black-box query-based and search-based attacks constitute a third methodology, applicable when the attacker lacks gradient access to the target model. These approaches include genetic algorithms and particle swarm optimization techniques that iteratively evolve candidate perturbations based solely on the binary or confidence-score feedback returned by querying the target IDS, as well as substitute-model approaches in which the attacker trains a local surrogate classifier on a dataset assumed to be similar to the target's training distribution, computes white-box adversarial examples against the surrogate, and transfers these examples to the black-box target, relying on the empirically observed transferability of adversarial perturbations across differently trained models [11], [15]. While black-box attacks generally require substantially more queries and achieve somewhat lower success rates than white-box counterparts, reported success rates in the range of fifty to seventy percent against commercial-grade or ensemble-based IDS architectures demonstrate that the absence of direct model access does not provide strong practical protection on its own [11].

A further methodology of particular relevance to network security, distinct from the general adversarial machine learning literature, is the mimicry attack, in which an attacker crafts malicious traffic engineered to closely resemble the statistical distribution of legitimate traffic observed in the target network, exploiting the natural overlap between benign and malicious feature distributions rather than relying on an explicit optimization procedure against a specific model. Mimicry attacks are notable because they do not require any knowledge of the target classifier whatsoever, relying instead on general knowledge of typical benign traffic patterns, and have been shown to be effective against anomaly-based detectors that rely heavily on statistical deviation from an established baseline of normal behavior [7]. Taken together, the reviewed methodologies demonstrate that adversarial threats against IDS are neither limited to a narrow set of idealized white-box conditions nor dependent on unrealistic attacker capabilities, but instead span a spectrum of practically achievable techniques applicable under varying degrees of attacker knowledge and access.

VI. DEFENSE AND MITIGATION STRATEGIES

The defense literature for adversarial machine learning in intrusion detection draws substantially from techniques originally developed in the computer vision robustness community, adapted where necessary to respect the problem-space constraints discussed earlier. Adversarial training remains the most widely adopted and empirically validated defense strategy, involving the augmentation of the training dataset with adversarial examples generated during training, typically through an iterative procedure such as PGD, so that the model learns to correctly classify both clean and perturbed inputs [17]. Madry et al. framed adversarial training as a min-max optimization problem in which the inner maximization finds the worst-case perturbation within an allowed budget and the outer minimization updates model parameters to minimize loss against that worst case, a formulation that has become the standard theoretical framing for robust training procedures [17]. Applied to IDS classifiers,

10.48047/jocaaa.2022.30.02.76

adversarial training has been shown to substantially narrow, though not eliminate, the gap between clean and adversarial accuracy, at the cost of increased training time and, in some studies, a modest reduction in clean-data accuracy relative to a non-adversarially trained baseline [10].

Defensive distillation, an earlier proposed defense, trains a secondary model to mimic the softened output probabilities of an initially trained model at an elevated temperature, with the intention of smoothing the decision surface and increasing the perturbation magnitude required to induce misclassification [22]. While early evaluations reported promising robustness gains, subsequent work by Carlini and Wagner demonstrated that defensive distillation, along with several other gradient-masking based defenses, could be defeated by attacks specifically designed to circumvent obfuscated or numerically unstable gradients, rather than genuinely increasing robustness [18], [23]. This finding underscores a broader methodological lesson that has become central to the adversarial robustness literature: a defense that merely obstructs gradient-based optimization without altering the underlying decision boundary in a fundamentally robust way tends to provide a false sense of security and can typically be circumvented once an adaptive attacker becomes aware of the specific obfuscation mechanism in use [23].

Ensemble and moving-target defenses seek robustness through diversity rather than through hardening any single model. Ensemble approaches combine predictions from multiple classifiers, potentially trained on different feature subsets or algorithmic families, under the reasoning that an adversarial perturbation effective against one architecture is unlikely to simultaneously fool all ensemble members, particularly when the ensemble is not fully known to the attacker. Moving-target defenses extend this idea temporally, periodically switching between models, feature subsets, or decision thresholds so that any adversarial example crafted by an attacker against the currently deployed configuration has a limited window of effectiveness before the underlying configuration changes, as proposed by Han et al. specifically for network intrusion detection contexts [16]. Both approaches introduce operational complexity and computational overhead that must be weighed against the real-time throughput requirements typical of production IDS deployments.

Feature squeezing and input transformation defenses attempt to reduce the sensitivity of a model to small perturbations by reducing the effective input space available to an attacker, for example by quantizing continuous features into coarser bins or applying smoothing transformations prior to classification, on the reasoning that an adversarial perturbation small enough to remain imperceptible or difficult to detect will often be eliminated by the transformation, while legitimate traffic characteristics survive the transformation largely intact. Detect-and-reject approaches take a complementary strategy, training a secondary classifier or statistical test specifically to identify inputs that exhibit characteristics consistent with adversarial manipulation, such as anomalously placed values near a decision boundary or statistical inconsistencies introduced by the perturbation process, and routing flagged inputs to a separate analysis pipeline rather than relying solely on the primary classifier's output.

Certified robustness techniques, including randomized smoothing, provide probabilistic guarantees that a classifier's prediction will remain unchanged for any perturbation within a specified bound, offering a theoretically stronger notion of robustness than empirical defenses that are only evaluated against a fixed set of known attacks. However, certified defenses typically impose substantial computational overhead at both training and inference time and have seen comparatively limited adoption in the network intrusion detection literature relative to the computer vision domain, in part due to the strict latency requirements of real-time traffic classification. Finally, several authors have proposed incorporating domain-aware feature

10.48047/jocaaa.2022.30.02.76

engineering directly into the defense strategy, deliberately selecting or transforming features so that any modification an attacker could plausibly achieve within the problem-space constraints discussed in Section III corresponds to only a small, bounded change in the feature-space representation, thereby narrowing the gap between problem-space feasibility and feature-space vulnerability that many attacks otherwise exploit [12], [19].

VII. EXPERIMENTAL EVALUATION FRAMEWORK AND DISCUSSION

A rigorous evaluation methodology is essential to assess both the severity of adversarial threats and the effectiveness of proposed defenses, yet the literature reveals considerable inconsistency in the datasets, metrics, and threat models adopted across studies. The most commonly used benchmark datasets are NSL-KDD, a refined version of the older KDD Cup 1999 dataset addressing several of its known redundancy and class-imbalance issues [24]; CICIDS2017, which provides more contemporary traffic captures including a broader range of modern attack categories such as brute force, denial of service, and infiltration; and UNSW-NB15, which combines synthetic modern attack behaviors with real background traffic and has become a standard benchmark for evaluating both detection accuracy and adversarial robustness [25]. Each dataset carries known limitations, including potential simulation artifacts that may not fully represent real production network conditions, a concern that becomes particularly salient when evaluating adversarial robustness, since an attack that appears highly effective on a dataset with unrealistic feature distributions may not generalize to genuine network deployments.

Standard evaluation metrics in this literature include clean accuracy, measuring detection performance on unperturbed test data; adversarial accuracy, measuring detection performance on adversarially perturbed test samples; and attack success rate, the complementary metric measuring the proportion of adversarial samples that successfully evade detection. Some studies additionally report the perturbation budget required to achieve a given attack success rate, expressed as an L-infinity or L2 norm bound, or as the number of features modified in sparse attacks such as JSMA, providing a proxy for how easily an attacker could realistically achieve the reported evasion rate. Comparing across the literature reviewed in Sections V and VI, several consistent patterns emerge: white-box gradient attacks tend to achieve the highest attack success rates but require the least realistic attacker knowledge assumption; GAN-based approaches achieve strong success rates with a more favorable computational profile for generating large volumes of adversarial traffic; black-box approaches achieve meaningfully lower but still practically significant success rates; and adversarial training consistently narrows, without fully closing, the gap between clean and adversarial accuracy across nearly all studies that report both figures [9], [10], [11], [17].

A recurring methodological concern, first articulated forcefully in the computer vision robustness literature but equally applicable to network intrusion detection, is the risk of evaluating a proposed defense only against the specific attacks known at the time of its publication, rather than against an adaptive attacker aware of the defense mechanism itself. Carlini and Wagner's demonstration that defensive distillation could be defeated once the specific obfuscation mechanism was understood exemplifies this concern directly [18], [23], and similar adaptive re-evaluation has been comparatively rare in the network intrusion detection literature, where many proposed defenses are validated against only the attack methods available at publication time. This gap suggests that reported robustness improvements for IDS defenses should be interpreted cautiously until subjected to adaptive evaluation analogous to the standards that have become customary in the broader adversarial machine learning research community.

10.48047/jocaaa.2022.30.02.76

Beyond raw detection metrics, practical deployment considerations weigh heavily on which defenses are viable in production IDS environments. Real-time network monitoring typically requires per-flow classification within milliseconds to support inline blocking decisions, imposing strict latency budgets that many computationally intensive defenses, such as certified robustness techniques or large ensembles, struggle to satisfy without specialized hardware acceleration. This tension between defensive robustness and operational latency represents one of the most consequential practical trade-offs distinguishing adversarial robustness research in network security from analogous work in domains such as image classification, where inference latency is typically a secondary concern relative to detection accuracy.

VIII. OPEN CHALLENGES AND FUTURE RESEARCH DIRECTIONS

Despite substantial progress documented in the surveyed literature, several open challenges remain insufficiently addressed and represent promising directions for future research. First, the absence of a standardized, widely adopted benchmark suite for adversarial robustness evaluation in network intrusion detection, comparable to benchmarks such as RobustBench in the computer vision domain, makes it difficult to compare reported robustness figures across studies that differ in dataset, feature set, attacker capability assumptions, and evaluation metric, and the field would benefit considerably from a community-agreed evaluation protocol that explicitly incorporates problem-space feasibility constraints.

Second, the tension between defense computational overhead and the real-time throughput demands of production intrusion detection deployments remains largely unresolved; while adversarial training and ensemble methods demonstrate measurable robustness improvements in academic evaluations, their practical deployment in high-throughput enterprise or carrier-grade networks, where per-packet or per-flow processing budgets are measured in microseconds, has received comparatively little attention, and lightweight or hardware-accelerated variants of these defenses represent an important applied research direction.

Third, most existing work has focused on evasion attacks against traffic classified using conventional flow-based features, while comparatively less attention has been devoted to adversarial vulnerabilities in encrypted traffic classification systems, which increasingly rely on side-channel features such as packet timing and size sequences rather than payload content; because these side-channel features are themselves susceptible to adversarial manipulation through techniques such as deliberate packet padding or timing perturbation, adversarial robustness in this emerging and increasingly important detection paradigm warrants dedicated investigation.

Fourth, the increasing use of continuously retrained or online-learning IDS models to adapt to evolving threats amplifies the practical significance of poisoning attacks relative to the comparatively better-studied evasion category, yet defenses specifically designed to detect or mitigate poisoning in the network security context, as opposed to the broader machine learning poisoning literature, remain comparatively underdeveloped. Finally, the emergence of large-scale foundation models and language-model-based approaches for network traffic analysis and threat intelligence introduces a new and largely unexplored attack surface, including the potential for prompt-injection-style manipulation of model-assisted security operations workflows, an area that is likely to grow in importance as such models become more deeply integrated into security operations centers.

IX. CONCLUSION

10.48047/jocaaa.2022.30.02.76

Machine learning and deep learning have substantially advanced the capability of intrusion detection systems to identify novel and previously unseen network attacks, but this advancement carries an inherent and well-documented cost: the same statistical learning mechanisms that enable generalization also create exploitable vulnerabilities to adversarial manipulation. This paper has presented a structured review of the adversarial threat landscape confronting learning-based IDS, organizing attacks into a taxonomy spanning evasion, poisoning, model extraction, privacy inference, and reconnaissance categories, and has surveyed the principal attack methodologies and corresponding defense strategies reported in the literature. The evidence reviewed indicates that reported clean-data detection accuracy alone provides limited assurance of practical resilience, that black-box attacks remain a credible threat even without direct model access, and that many proposed defenses, while measurably improving robustness against known attacks, have not been subjected to the rigorous adaptive evaluation standards that have become customary in the broader adversarial machine learning community. Addressing the open challenges identified in this paper, particularly the development of standardized, problem-space-aware evaluation benchmarks and the design of computationally efficient defenses suitable for real-time deployment, will be essential to closing the gap between academically demonstrated robustness and genuine operational resilience against motivated, adaptive adversaries in production network environments.

REFERENCES

[1] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: Methods, systems and tools," IEEE Communications Surveys & Tutorials, vol. 16, no. 1, pp. 303-336, 2014.

[2] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153-1176, 2016.

[3] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," IEEE Access, vol. 7, pp. 41525-41550, 2019.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in Proc. Int. Conf. Learning Representations (ICLR), 2014.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[6] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," Pattern Recognition, vol. 84, pp. 317-331, 2018.

[7] M. Rigaki, "Adversarial deep learning against intrusion detection classifiers," M.S. thesis, Luleå University of Technology, 2017.

[8] K. Yang, J. Liu, C. Zhang, and Y. Fang, "Adversarial examples against the deep learning based network intrusion detection systems," in Proc. IEEE Military Communications Conf. (MILCOM), 2018, pp. 559-564.

[9] M. Usama, J. Qadir, and A. Al-Fuqaha, "Adversarial attacks on cognitive self-organizing networks: The challenge and the way forward," in Proc. IEEE Local Computer Networks Workshops (LCN Workshops), 2018, pp. 90-97.

10.48047/jocaaa.2022.30.02.76

[10] G. Apruzzese, M. Colajanni, L. Ferretti, and M. Marchetti, "Addressing adversarial attacks against security systems based on machine learning," in Proc. Int. Conf. Cyber Conflict (CyCon), 2019, pp. 1-18.

[11] E. Alhajjar, P. Maxwell, and N. Bastian, "Adversarial machine learning in network intrusion detection systems," Expert Systems with Applications, vol. 186, 115782, 2021.

[12] R. Sheatsley, N. Papernot, M. Weisman, G. Verma, and P. McDaniel, "Adversarial examples for network intrusion detection systems," Journal of Computer Security, vol. 30, no. 5, pp. 727-752, 2021.

[13] O. Ibitoye, R. Abou-Khamis, A. Matrawy, and M. O. Shafiq, "The threat of adversarial attacks on machine learning in network security -- A survey," arXiv preprint arXiv:1911.02621, 2019.

[14] E. Anthi, L. Williams, M. Rhode, P. Burnap, and A. Wedgbury, "Adversarial attacks on machine learning cybersecurity defences in industrial control systems," Journal of Information Security and Applications, vol. 58, 102717, 2021.

[15] I. Debicha, T. Debatty, J.-M. Dricot, and W. Mees, "Adversarial training for deep learning-based intrusion detection systems," in Proc. Int. Conf. Emerging Security Information, Systems and Technologies (SECURWARE), 2021.

[16] Y. Han, B. I. P. Rubinstein, T. Abraham, T. Alpcan, O. De Vel, S. Erfani, D. Hubczenko, C. Leckie, and P. Montague, "Reinforcement learning for autonomous defence in software-defined networking," in Proc. Int. Conf. Decision and Game Theory for Security (GameSec), 2018, pp. 145-165.

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proc. Int. Conf. Learning Representations (ICLR), 2018.

[18] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in Proc. IEEE Symposium on Security and Privacy (S&P), 2017, pp. 39-57.

[19] F. Pierazzi, F. Pendlebury, J. Cortellazzi, and L. Cavallaro, "Intriguing properties of adversarial ML attacks in the problem space," in Proc. IEEE Symposium on Security and Privacy (S&P), 2020, pp. 1332-1349.

[20] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in Proc. IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.

[21] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. Int. Conf. Information Systems Security and Privacy (ICISSP), 2018, pp. 108-116.

[22] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in Proc. NIPS Deep Learning Workshop, 2015.

[23] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in Proc. Int. Conf. Machine Learning (ICML), 2018, pp. 274-283.

[24] L. Dhanabal and S. P. Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," Int. Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 6, pp. 446-452, 2015.

[25] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," in Proc. Military Communications and Information Systems Conf. (MilCIS), 2015, pp. 1-6.