

Machine Learning for Intrusion Detection and Prevention Systems: A Review of Approaches and Challenges

Khayal Farzaliyev^{1,*}

¹ National Aviation Academy, Azerbaijan, Clearwater 1000, Azerbaijan

Email: ¹ khayal.farzaliyev@naa.edu.az

*Corresponding Author

Abstract—Purpose – Machine learning (ML) and deep learning (DL) have emerged as the dominant paradigm for intrusion detection and prevention systems (IDPS), and detection accuracies reported for benchmark datasets such as CIC-IDS2017, UNSW-NB15, and NSL-KDD now consistently surpass 99%. To situate this progress, the present review examines the main families of ML-based IDPS mechanisms, encompassing classical supervised learning, deep and hybrid architectures, unsupervised and semi-supervised anomaly detection, and ensemble and federated approaches.

Design/methodology/approach – Representative and highly cited studies were found in the major scholarly databases and arranged according to detection strategy, deployment setting, and learning paradigm. The recurring challenges that restrict their operational deployment were consolidated, namely class imbalance, weak cross-dataset generalization, adversarial fragility, limited interpretability, and computational cost.

Findings – The synthesis exposes a structural bias in the literature, in that the vast majority of studies assess detection quality on full, post-hoc flow records under the implicit assumption of unbounded inspection capacity. In contrast, there is relatively little research on early flow classification, which involves making decisions based on the initial packets of a flow before it concludes. This capability is crucial for inline prevention, reducing latency, and optimizing resource-aware scheduling.

Originality – The review shows that the disparity between offline accuracy and online, early, and resource-limited decision-making constitutes the most consequential open problem for practical IDPS, and it outlines a corresponding research agenda that reframes detection as an early, resource-aware decision problem rather than a purely accuracy-driven one.

Keywords—Intrusion detection, Machine learning, Deep learning, Early flow classification, Network security

I. INTRODUCTION

Network intrusion detection and prevention systems (IDPS) constitute a core defensive control for modern infrastructure, in that detection systems (IDS) monitor traffic and raise alerts on suspicious activity, whereas prevention

systems (IPS) operate inline and can actively block or drop malicious flows. Both were historically signature-based, matching traffic against databases of known attack patterns, and although such approaches are precise for known threats, they cannot recognize novel or zero-day attacks and require constant signature updates [1]. This limitation motivated a decisive shift toward anomaly-based detection driven by machine learning (ML) and, more recently, by deep learning (DL), each of which learns patterns of normal and malicious behavior directly from data and can, in principle, generalize to previously unseen attacks [1, 2].

The past decade has witnessed an enormous body of ML-based IDPS research, and on widely used benchmarks the reported accuracies are now close to saturation, with hybrid deep models reaching approximately 99% on standard datasets [3]. On the surface, these results would imply that the intrusion-classification problem has been largely solved; nevertheless, operational deployments continue to struggle. This indicates a persistent misalignment between the objectives that the literature optimizes and the practical requirements of deployed systems [4].

This review pursues three goals. The first is to organize the ML-based IDPS landscape into coherent methodological families and to summarize representative work within each of them. The second objective is to consolidate the common challenges that recur across these families and that bound their practical effectiveness, namely class imbalance, weak generalization, adversarial fragility, limited interpretability, and computational cost. The third, and the most significant, is to examine a fundamental gap that the accuracy-centric perspective has obscured. More specifically, the lack of studies on early flow classification, which involves producing a reliable decision from the first few packets of a flow before that flow has completed, and which is precisely the regime in which an inline IPS is required to operate.



Received: 14-8-2026

Revised: 10-9-2026

Published: 31-12-2026

II. MATERIALS AND METHODS

This work is structured as a narrative review of the peer-reviewed literature on ML-based and DL-based network intrusion detection and prevention. Representative and highly cited studies published between 2015 and 2025 were identified from the major scholarly databases, including IEEE Xplore, SpringerLink, and preprint repositories, using combinations of the search terms intrusion detection, intrusion prevention, machine learning, deep learning, and the names of the standard benchmark datasets. Studies were selected if they proposed or rigorously evaluated an ML-based detection mechanism on a recognized benchmark, and foundational surveys were included for historical grounding [1, 5, 6].

The retrieved studies were organized along three axes. According to their detection strategy, IDPS mechanisms are classified as signature-based, anomaly-based, or hybrid; according to their deployment, they are classified as network-based (NIDS) or host-based (HIDS), and according to their learning paradigm, they are classified as supervised, unsupervised, semi-supervised, ensemble, or federated [1]. ML-based approaches primarily focus on the anomaly-based NIDS setting, in which detection is modeled either as a supervised classification problem or as an unsupervised outlier-detection problem.

Benchmark datasets have significantly influenced the direction of the field, and the studies reviewed in this work rely almost exclusively on them. KDD CUP 99 and its cleaned successor NSL-KDD dominated early work [7], whereas the more realistic CIC-IDS2017 [8, 9], together with CSE-CIC-IDS2018 [10] and UNSW-NB15 [11], have since become standard. CIC-IDS2017 alone consists of roughly 2.8 million labeled flows captured over a five-day period, represented by 78 flow features, and it is also heavily imbalanced toward benign traffic [8]. These dataset characteristics, and especially the marked imbalance and the reliance on full flow records, implicitly shape the methods that are subsequently built on them. This issue is discussed further in the Discussion section.

III. RESULTS

A. Classical supervised learning approaches

The foundational surveys of the field, and most prominently the survey by Buczak and Guven [1], catalogued the adoption of decision trees, support vector machines, naïve Bayes, k-nearest neighbors, random forests, and ensemble methods to intrusion detection, thereby establishing many of the evaluation conventions that remain in use today. Traditional supervised models continue to serve as strong baselines, given that ensemble learners are frequently recognized as one of the most reliable performers [12]. From a practical perspective, tree-based and related models

continue to be favored for computationally efficient detection that is suitable for resource-constrained and real-time environments [13]. Due to their ability to train quickly, operate on tabular flow features without heavy preprocessing, and provide low-latency inference, such models are strong candidates for the lightweight, early-decision role highlighted later, and yet the existing studies largely deploy them on complete flow records rather than on partial ones.

B. Deep and hybrid architectures

Deep learning entered the IDPS domain primarily through feature learning, as demonstrated by Javaid et al. [14], who employed self-taught learning with a sparse autoencoder on NSL-KDD, which showed that unsupervised pretraining could extract discriminative patterns for intrusion classification. Convolutional neural networks were subsequently introduced to learn spatial feature interactions [15], and deep neural networks more generally were shown to be effective on benchmark traffic [16]. Recurrent architectures and long short-term memory networks were utilized to capture the temporal structure of packet sequences and thereby enhance the detection of stealthy attacks [2]. Building on this approach, deep recurrent models were later extended to resource-constrained IoT settings [17].

The current mainstream consists of hybrid architectures that compose complementary inductive biases, and the reported results are correspondingly strong, as exemplified by transformer and convolutional hybrids that incorporate resampling-based imbalance mitigation and report high accuracy on recent benchmarks [3]. Systematic reviews of DL-based IDS document this proliferation while simultaneously noting a number of persistent methodological weaknesses [5, 6]. A recurring trend in the literature is a gradual increase in benchmark accuracy that is accompanied by rising model complexity and inference cost. Critically, this trade-off becomes problematic for inline prevention, in which the per-flow latency budget is severely limited.

C. Unsupervised and semi-supervised detection

Due to the lack of sufficient labeled attack data and the inherently unlabeled nature of novel attacks, unsupervised and semi-supervised methods are especially suitable for zero-day detection. Deep autoencoders employed as one-class classifiers trained exclusively on benign traffic and identifying high reconstruction errors as anomalies have undergone extensive evaluation. Furthermore, they have been combined with support vector machines to create effective unsupervised detectors [18]. The deep one-class classification framework extended this concept and provided a principled objective for anomaly detection with neural networks [19]. These methods partially address the problems of label scarcity and novelty. However, they still follow the same evaluation protocol as the supervised literature, as they

are typically assessed on complete flows and in an offline setting.

D. Ensembles, federated learning, and imbalance handling

A distinct methodological cluster has emerged due to two practical issues. The major limitation is class imbalance, since when benign traffic dominates, naïve models often report high accuracy despite failing to detect rare attacks. Standard solutions to this comprise resampling, among which the synthetic minority over-sampling technique is the most prominent instance [20], along with cost-sensitive learning and ensemble techniques. The second concern is that of generalization and reliability, which ensemble methods mitigate by combining heterogeneous learners into more robust detectors [12], and which surveys of big-data intrusion detection examine across successive benchmark generations [10]. The same studies that report strong within-dataset ensemble performance, however, also surface the most uncomfortable finding of the field. In particular, performance frequently collapses when models are transferred across datasets.

E. Cross-cutting challenges

1. Class imbalance

Imbalance is the most frequently reported barrier, because benign-dominated data tend to produce inflated accuracy estimates while simultaneously suppressing recall on the minority attack classes, and although oversampling and cost-sensitive learning mitigate this problem, they do not eliminate it. A remaining challenge is that aggressive resampling can introduce overfitting to synthetic minority points [20]. Imbalance moreover interacts with evaluation itself, in that accuracy constitutes a poor headline metric under these conditions, and yet it continues to be widely reported.

2. Weak cross-dataset generalization

Models trained on a single dataset often generalize poorly to other datasets, and systematic reviews indicate that detection performance degrades significantly once a model is evaluated outside its training distribution, a finding emphasizing the need for comprehensive cross-dataset validation [5]. This fragility implies that high in-dataset accuracy largely reflects dataset-specific artifacts rather than any transferable knowledge of the underlying attacks.

3. Adversarial fragility

ML-based detectors are susceptible to adversarially perturbed traffic that is crafted specifically to evade classification. Recent reviews further indicate that robustness remains an open problem in DL-based detection [6]. In the case of an inline IPS, which constitutes an adversary's direct target, this fragility is far from academic, since it represents a primary attack surface that benchmark accuracy alone does not measure.

4. Interpretability and trust

High-capacity deep models are inherently opaque, which complicates both analyst triage and incident response. Although explainability has been proposed as a means of improving trust without materially sacrificing accuracy, explainable IDPS nevertheless remains not yet widely deployed [6].

5. Computational cost and deployability

Rising accuracy has involved a trade-off with increased complexity, since deep and hybrid models impose inference latency and memory costs that strain deployment on edge, IoT, and high-throughput inline settings. In these settings, lightweight models are explicitly favored [13, 17]. Systematic surveys of the deployment of ML for traffic classification further emphasize that laboratory accuracy seldom transfers to operational constraints [4]. Therefore, cost is not a secondary concern but rather a first-order determinant of whether a given method can serve as an IPS at all.

F. Comparative summary of representative approaches

Table I provides an overview of representative work across the families discussed above and contrasts the methodological focus of each one. Specifically, it considers two properties that are central to inline prevention, namely whether the approach addresses class imbalance and whether its decision is reached on complete or on partial flows. The final column renders the central observation of this review concrete, since almost every high-accuracy method operates exclusively on complete flows.

Table 1. Representative ML-based IDPS approaches and their reliance on complete flow records

Study	Approach / model family	Dataset(s)	Imbalance handling	Flow input
Buczak & Guven [1]	Survey of classical ML/DM methods	Multiple	Discussed	Complete
Javaid et al. [14]	Self-taught learning; sparse autoencoder	NSL-KDD	No	Complete
Yin et al. [2]	LSTM / recurrent sequence model	NSL-KDD	No	Complete
Al-Qatf et al. [18]	Sparse autoencoder + SVM	NSL-KDD	No	Complete
Ruff et al. [19]	Deep one-class classification	Multiple	One-class	Complete
Almiani et al. [17]	Deep recurrent network (IoT)	Benchmark	No	Complete
Kamil & Mohammed [15]	Convolutional neural network	Benchmark	No	Complete

Study	Approach / model family	Dataset(s)	Imbalance handling	Flow input
Hossain & Islam [12]	Ensemble machine learning	Benchmark	Ensemble	Complete
Kamal & Mashaly [3]	Transformer-CNN + resampling	Recent benchmarks	Resampling	Complete
Lansky et al. [5]	Systematic review of DL-based IDS	Multiple	Reviewed	Complete
Pacheco et al. [4]	Deployment survey (incomplete flows)	Multiple	—	Partial

IV. DISCUSSION

A. The overlooked problem of early flow classification

The challenges consolidated in the Results are widely acknowledged. This section argues instead for a gap that is considerably less visible precisely because it is structural, in that it is embedded in the manner in which the field frames the problem rather than in any single method. Most studies evaluate detection on complete flow records, so that features such as the total flow duration, the total number of bytes, and the packet-length statistics are computed only after a flow has finished [8]. Thus, the classifier consequently judges the flow only after the relevant activity has already occurred. Although this arrangement is entirely appropriate for an offline IDS that produces alerts for subsequent analysis, it is fundamentally misaligned with an inline IPS, which must decide whether to permit or block a flow while that flow is still in progress and ideally from its first few packets, because any verdict that arrives after the flow has completed can no longer prevent the associated harm.

This mismatch entails three consequences that the accuracy-centric literature does not capture. The first consequence concerns latency, in that complete-flow features are simply unavailable at the moment of decision for a live flow, meaning that the reported accuracies systematically overstate what an inline system can achieve at an early stage. The second consequence concerns resource allocation, since an inline engine operating under traffic load must choose which flows to inspect first and which to drop when capacity is exceeded. An early suspicion estimate is precisely the signal required to schedule such inspection, yet the literature rarely produces one. The third consequence concerns prevention semantics, because blocking is meaningful only before damage has occurred, which requires an early decision by construction.

The evidence that this problem is genuinely under-explored is itself sparse. A small number of recent studies have begun to examine the deployment challenges that arise when detection must act upon incomplete flows [4], yet work that

systematically characterizes the accuracy–earliness trade-off and integrates early classification with inline prevention remains scarce compared with the extensive complete-flow literature. The dominant incentive structure, which rewards the maximization of benchmark accuracy on complete flows, appears to steer researchers away from the harder and more operationally relevant early-decision problem.

B. Research gaps and future directions

From the foregoing analysis, four directions are distilled as the most consequential for future work.

- The first direction concerns early flow classification, and it calls for a systematic characterization of the trade-off between accuracy and earliness, in the sense of establishing how reliably attacks can be identified from the first few packets and how detection quality degrades as the available evidence is reduced. This is the central under-explored problem, and it is at the same time the prerequisite for genuine inline prevention.
- The second direction concerns resource-aware inspection and scheduling, requiring the field to move beyond the implicit assumption of unbounded capacity by using early suspicion estimates to prioritize which flows are inspected first and which are shed under load. This reframes detection as a scheduling problem rather than a pure classification problem.
- The third direction concerns robust cross-dataset and adversarial evaluation, requiring cross-dataset validation as standard practice and adversarial robustness reporting alongside clean accuracy. This ensures that the performance reported reflects transferable and attack-resistant capability rather than a dataset-specific fit [5, 6].
- The fourth direction concerns lightweight and interpretable models for deployment, and it involves prioritizing models whose inference cost and transparency are suited to edge, IoT, and high-throughput inline settings, while reporting latency and memory as first-class metrics rather than as afterthoughts [4, 17].

V. CONCLUSION

Research on ML-based and DL-based IDPS has achieved near-saturated accuracy on the standard benchmarks. The present review has organized this progress across the classical, deep, hybrid, unsupervised, ensemble, and federated families, while consolidating the challenges of imbalance, generalization, adversarial fragility, interpretability, and cost. The most important limitation of the field is nevertheless structural, as it overwhelmingly evaluates complete flows in an offline setting under an

implicit assumption of unbounded inspection capacity. Comparatively little work addresses the reliable early decision required for inline prevention or connects such decisions to resource-aware inspection under load. This review argues that addressing this gap, rather than pursuing gains on saturated benchmarks, is where the next meaningful advances in practical IDPS are likely to emerge.

REFERENCES

- [1] Buczak AL, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Commun Surv Tutor.* 2016; 18(2): 1153-1176. doi: 10.1109/COMST.2015.2494502
- [2] Yin C, Zhu Y, Fei J, He X. A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access.* 2017; 5: 21954-21961. doi: 10.1109/ACCESS.2017.2762418
- [3] Kamal H, Mashaly M. Advanced hybrid transformer-CNN deep learning model for effective intrusion detection systems with class imbalance mitigation using resampling techniques. *Future Internet.* 2024; 16(12): 481. doi: 10.3390/fi16120481
- [4] Pacheco F, Exposito E, Gineste M, Baudoin C, Aguilar J. Towards the deployment of machine learning solutions in network traffic classification: a systematic survey. *IEEE Commun Surv Tutor.* 2019; 21(2): 1988-2014. doi: 10.1109/COMST.2018.2883147
- [5] Lansky J, Ali S, Mohammadi M, Majeed MK, Karim SHT, Rashidi S, et al. Deep learning-based intrusion detection systems: a systematic review. *IEEE Access.* 2021; 9: 101574-101599. doi: 10.1109/ACCESS.2021.3097247
- [6] Wu Y, Zou B, Cao Y. Current status, challenges and future trends of deep learning-based intrusion detection models. *J Imaging.* 2024; 10(10): 254. doi: 10.3390/jimaging10100254
- [7] Tavallaei M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the KDD CUP 99 data set. In: 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA); 2009. p. 1-6. doi: 10.1109/CISDA.2009.5356528
- [8] Sharafaldin I, Lashkari AH, Ghorbani AA. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP); 2018. p. 108-116. doi: 10.5220/0006639801080116
- [9] Stiawan D, Idris MYB, Bamhdi AM, Budiarto R. CICIDS-2017 dataset feature analysis with information gain for anomaly detection. *IEEE Access.* 2020; 8: 132911-132921. doi: 10.1109/ACCESS.2020.3009843
- [10] Leevy JL, Khoshgoftaar TM. A survey and analysis of intrusion detection models based on CSE-CIC-IDS2018 Big Data. *J Big Data.* 2020; 7: 104. doi: 10.1186/s40537-020-00382-x
- [11] Moustafa N, Slay J. UNSW-NB15: a comprehensive data set for network intrusion detection systems. In: 2015 Military Communications and Information Systems Conference (MilCIS); 2015. p. 1-6. doi: 10.1109/MilCIS.2015.7348942
- [12] Hossain MA, Islam MS. Ensuring network security with a robust intrusion detection system using ensemble-based machine learning. *Array.* 2023; 19: 100306. doi: 10.1016/j.array.2023.100306
- [13] Yang L, Li J, Yin L, Sun Z, Zhao Y, Li Z. Real-time intrusion detection in wireless network: a deep learning-based intelligent mechanism. *IEEE Access.* 2020; 8: 170128-170139. doi: 10.1109/ACCESS.2020.3019973
- [14] Javaid A, Niyaz Q, Sun W, Alam M. A deep learning approach for network intrusion detection system. In: Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (BICT); 2016. p. 21-26. doi: 10.4108/eai.3-12-2015.2262516
- [15] Kamil WF, Mohammed IJ. Deep learning model for intrusion detection system utilizing convolution neural network. *Open Eng.* 2023; 13(1): 20220403. doi: 10.1515/eng-2022-0403
- [16] Maithem M, Al-Sultany GA. Network intrusion detection system using deep neural networks. *J Phys Conf Ser.* 2021; 1804(1): 012138. doi: 10.1088/1742-6596/1804/1/012138
- [17] Almiani M, AbuGhazleh A, Al-Rahayfeh A, Atiewi S, Razaque A. Deep recurrent neural network for IoT intrusion detection system. *Simul Model Pract Theory.* 2020; 101: 102031. doi: 10.1016/j.simpat.2019.102031
- [18] Al-Qatf M, Lasheng Y, Al-Habib M, Al-Sabahi K. Deep learning approach combining sparse autoencoder with SVM for network intrusion detection. *IEEE Access.* 2018; 6: 52843-52856. doi: 10.1109/ACCESS.2018.2869577
- [19] Ruff L, Vandermeulen R, Goernitz N, Deecke L, Siddiqui SA, Binder A, et al. Deep one-class classification. In: Proceedings of the 35th International Conference on Machine Learning (ICML); 2018. PMLR 80: 4393-4402.
- [20] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* 2002; 16: 321-357. doi: 10.1613/jair.953