INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

MACHINE LEARNING FOR NETWORK INTRUSION DETECTION: A COMPREHENSIVE REVIEW OF METHODS, BENCHMARKS, AND OPEN CHALLENGES

Prisha Senthil[1], Sabarish Kumaresan[2], Dr. P. Janarthanan[3]
[1,2,3]Department of Computer Science and Engineering, Sri Venkateswara College of Engineering, Sriperumbudur, India

Abstract — Network Intrusion Detection Systems (NIDS) form one of the key foundations of cybersecurity. Explosive expansion of network traffic, increasing popularity of encryption and new types of attacks make signature-based and rule-based detection techniques obsolete and inadequate for modern environments. In response to this challenge, machine learning (ML) becomes the leading approach in NIDS research, providing highly adaptive solutions far beyond simple pattern matching. In this paper we perform an extensive survey of machine learning-based NIDS: we analyze their most popular algorithmic classes, discuss benchmarking datasets used for evaluation, and highlight challenges preventing them from practical implementation. Classic supervised and ensemble approaches, deep learning models, multi-stage systems combining both ML and traditional detection techniques, and promising trends such as federated and explainable AI are covered. The current state of key challenges facing NIDS development – class imbalance, lack of up-to-date data, time constraints, and generalization limitations – is discussed and future research directions are outlined.

Keywords — Network Intrusion Detection, Machine Learning, Deep Learning, Ensemble Methods, Explainable AI,

1.INTRODUCTION

With the development of infrastructure in the world wide web network, it is possible to observe the emergence of the attack surface which is larger than any other before. The importance of the detection of malicious traffic for securing valuable data and maintaining high level of services and complying with the regulations can not be underestimated [1]. Traditionally, NIDS solutions based on the use of signatures and rules of detection have proven themselves as very effective in detection of known attacks but completely ineffective in detection of unknown exploits and malware, especially in case when the latter operate within TLS tunnel [2].

Machine learning techniques offer another perspective on network security: instead of looking for signatures of known attacks, the goal of such techniques is to generalize about previously unseen data on the basis of the data set of labeled attacks[3] . Machine learning techniques evolved through several stages from simple linear classifiers to ensembles and finally neural networks. Although there was much progress achieved in benchmarking and achieving high accuracy results, the gap between lab tests and practical application of such methods still exists [4].

2. TAXONOMY OF MACHINE LEARNING APPROACHES FOR NIDS

ML-based intrusion detection methods can be organized into four broad families: classical supervised learners, ensemble methods, deep learning architectures, and hybrid multi-stage pipelines. Each occupies a distinct point in the accuracy–latency–interpretability trade-off space.

2.1 CLASSICAL SUPERVISED METHODS

Initially, research on NIDS with machine learning techniques made use of SVMs, naive Bayes classifiers, k-NN methods, and decision trees. Although they provide good interpretability and fast inference, classical methods do not have enough expressiveness for non-linear high-dimensional feature spaces found in current network traffic data. Classical methods consistently yield lower recall on attacks that belong to the minority classes, such as U2R and R2L intrusions, where training samples are very rare [5].

2.2 ENSEMBLE METHODS

The introduction of ensemble approaches significantly improved baseline NIDS results. RF combines bootstrapped decision trees with random feature selection to create robust, low variance models capable of processing high-dimensional traffic features [6]. Gradient boosting ensembles, such as XGBoost and LightGBM, obtain state-of-the-art accuracy through iterative residuals updating and have shown excellent results in the context of NSL-KDD and UNSW-NB15 datasets [7]. Hozouri et al. discovered that higher variation in base learners of ensemble leads to fewer false positives in intrusion detection systems; this discovery became a motivation for further studies in multi-classifiers' designs [8]. The main drawback of ensemble approaches is that they tend to perform poorly in case of class imbalance: attack classes which are less than 1% in the training dataset

2.3 DEEP LEARNING ARCHITECTURES

CNNs and LSTM networks are widely utilized in NIDS applications. CNN models are effective in identifying spatial relationships between network traffic features, whereas LSTM networks are capable of learning temporal dependencies from sequential packet streams. Hybrid CNN-LSTM frameworks, which combine both spatial and temporal feature learning, have demonstrated better performance than single-model architectures on datasets such as CIC-IDS-2017 and CIC-IDS-2018 [10]. Attention-based mechanisms have further enhanced these models by assigning greater importance to highly discriminative features. Alam et al. reported that the Attention-CNN-LSTM approach achieved improved multi-class detection performance, particularly for complicated and infrequent attack categories [11]. Autoencoder-based methods have also gained attention for anomaly detection tasks by learning compact representations of normal network behavior and identifying deviations as malicious activity. Such approaches are especially useful for detecting zero-day attacks in situations where labeled attack data are limited or unavailable [12].

Although deep learning approaches provide superior detection accuracy, they require considerable computational resources and large-scale training datasets. Kikissagbe et al. observed that hybrid deep learning methods improved intrusion detection performance compared with standalone models; however, their computational overhead reduced suitability for real-time and latency-sensitive environments [13]. Balancing high detection accuracy with efficient real-time processing therefore remains a major challenge in modern NIDS research.

2.4 HYBRID MULTI-STAGE PIPELINES

Hybrid approaches tackle the problem of accuracy-latency trade-offs by breaking down classification into a sequence of stages having varying computational capacities. A light-weight primary classifier deals with the bulk amount of data, while the computationally more costly secondary classifier deals with only those few hard samples. The concept of confidence-based classification allows achieving this objective by retaining throughput while focusing computation on those samples which pose real challenges; i.e., the ones in which conventional single-step solutions would typically fail. The confidence-based approach has been successfully used for multi-classifier systems in vehicular communication networks, where Hossain et al. confirmed its usefulness for improving performance against various kinds of attacks [14].

3. BENCHMARK DATASETS

The choice of evaluation dataset strongly conditions reported performance figures and their generalizability. Table 1 summarizes the most widely used NIDS benchmarks, their principal characteristics, and known limitations.

Table 1. Summary of Widely Used NIDS Benchmark Datasets

Dataset	Year	Key Features	Known Limitations
NSL-KDD	2009	Improved KDD'99; 4 attack categories; widely cited	Aging traffic patterns; limited modern attack coverage
UNSW-NB15	2015	9 attack types; modern traffic mix; 49 features	Lab-generated; limited encrypted traffic
CIC-IDS-2017	2017	Realistic HTTP/HTTPS traffic; 15 attack types	Class imbalance up to 191,678:1 for rare attacks
CIC-IDS-2018	2018	Multi-day captures; 7 attack scenarios	Temporal redundancy; imbalanced classes
CIC-IoT-2023	2023	Large-scale IoT attack scenarios; real-time benchmark	Domain-specific; limited generalization outside IoT

The cross-datasets performance evaluation study of classifiers developed for use in cyber-attacks detection, conducted by Zoghi & Serpen, demonstrated that classifier systems with good performance on one particular dataset may show poor generalizability on another dataset.

4. PERSISTENT CHALLENGES

4.1 CLASS IMBALANCE

The imbalance between the majority benign traffic and the minority malicious activities constitutes the most common problem encountered in NIDS development work. Benign flows represent the vast majority in real-life and artificial datasets, whereas critical attacks, especially U2R and R2L categories, make up less than 1% of the data. Consequently, models trained on such skewed datasets are expected to produce substantial bias in favor of the majority, resulting in high accuracy values yet very low recall for the classes that pose serious threats. A number of methods have been proposed to address this issue. Oversampling involves synthesizing synthetic samples in the feature space of the minority class using interpolation, thus creating an objective way of rebalancing the dataset without ignoring the majority class [18]. Recent comparative studies on UNSW-NB15 and TON-IoT reveal that although classical oversampling techniques, e.g. SMOTE and ADASYN, produce consistent and interpretable gains, deep generative models, such as VAE and GANs, display better capabilities with higher variance depending on parameters [19]. On the other hand, algorithmic techniques aim at changing the objective function used during training to penalize minority-class errors [20].

4.2 REAL-TIME PROCESSING AND LATENCY CONSTRAINTS

Real operational NIDSs have to handle packets at wire speeds without any delay in detection that could enable attacks to be spread while waiting for their detection. Complex models capable of attaining higher benchmark accuracy, especially when we talk about multi-layer neural networks with millions of parameters, result in delays in inference that are too high to implement such systems operationally. This is the point where there is a trade-off between accuracy and latency. It implies that while simple classifiers reduce processing time per flow, they fail to provide discrimination ability due to poor accuracy and their incapability of dealing with uncommon classes of attacks .

4.3 ENCRYPTED TRAFFIC AND FEATURE AVAILABILITY

The widespread use of TLS 1.3 and HTTPS technology has made payload-based features impossible to access except through man-in-the-middle attacks, posing serious issues for privacy and potential lawbreaking in numerous countries. On the other hand, behavior-based and metadata-based features such as connection time, total bytes transferred per session, inter-arrival times, and protocol flags can be accessed even without examining

payload information, and have been successful enough to classify applications on well-known datasets. Nonetheless, attackers that know about detection techniques can manipulate behavior-based features, prompting further research into adversarial robustness.

4.4 GENERALIZATION ACROSS ENVIRONMENTS

Most evaluation studies on NIDS have been carried out using datasets generated in a single environment, and the models that result from them do not generalize well across organizations, networks, or even attack types. The diversities associated with cloud computing, the Internet of Things (IoT), and software-defined networking, all of which generate diverse traffic volumes, make single environment training even more difficult.Domain transfer learning and transfer learning generally seem like very promising avenues for research but have received less attention compared to related fields of ML.

5. EMERGING RESEARCH DIRECTIONS

5.1 FEDERATED LEARNING FOR PRIVACY-PRESERVING NIDS

Training a global model for detecting attacks in federated learning (FL) involves combining local model updates computed at distributed nodes in a network, without the need for centralization of raw traffic data, hence ensuring privacy of data. Federated learning NIDS have been found feasible in IoT and vehicle networks, with the recent inclusion of Byzantine resistant aggregation rules to resist attempts of manipulating the global model through adversarial client updates. One of the challenges associated with federated learning NIDS is the non-IID characteristics of the local traffic; different nodes see different traffic distributions.

5.2 EXPLAINABLE AI (XAI) FOR NIDS

The use of ML-based NIDS in a practical security environment demands that the security analyst trusts the decisions made by such models. It is impossible to audit, tweak, or justify decisions by black-box models that indicate malicious activity but lack interpretable reasoning. Methods of making NIDS algorithms explainable include SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). These approaches were used in NIDS workflows to determine the network characteristics that had the biggest impact on the decision [25]. Recently designed federated XAI-IDS models proved that the use of an explanation module did not significantly decrease accuracy rates while increasing analyst trust significantly. Interpretability was also helpful in anomaly detection on clients of federated models and debugging.

6. COMPARATIVE ANALYSIS OF REPRESENTATIVE WORKS

Table 2 summarizes representative NIDS studies across algorithmic families, highlighting the evaluation dataset, reported accuracy, and principal limitation of each approach.

Table 2. Comparative Summary of Representative ML-based NIDS Studies

Study	Method	Accuracy	Key Limitation
Haider et al. [13]	Hybrid DL + ML	98.72%	High computational overhead
Alashjaee et al. [11]	Attention-CNN-LSTM	99.1%	Training complexity; latency
Christy et al. [14]	Multi-stage Ensemble	98.94%	Limited to specific domain
Kikissagbe et al. [13]	SMOTE + Ensemble Classifier	98.4%	Sensitive to synthetic data quality

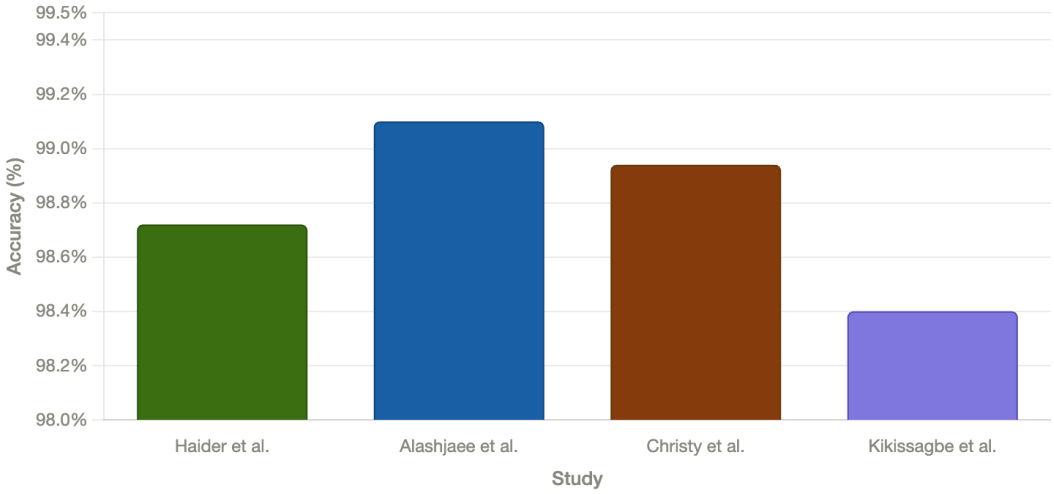

Figure 1. Accuracy comparison of representative ML-based NIDS studies

In summary, a common trend in this selection of representative papers is that techniques that focus on maximizing performance in terms of accuracy on a particular dataset will necessarily sacrifice performance in real-time, and techniques that perform well at the level of light-weight real-time detection fail at detecting attacks from minorities. It is impossible for a technique to be optimized along all four axes of performance at once.

7. DISCUSSION AND FUTURE DIRECTIONS

The history of ML-based NIDS research can thus be characterized by the successful establishment of powerful benchmarks coupled with a list of architectural weaknesses that are obscured under their numerical values. There are several promising lines of investigation in this context. For instance, there could be more emphasis on standardized multi-dataset evaluation methods that test NIDS for their ability to generalize across environments in addition to reporting the usual accuracy rates on one dataset. In relation to the second limitation discussed above, more attention should be paid to overcoming the class imbalance issue using approaches that go beyond SMOTE and focus on generating realistic minority-class traffic via conditional generative models. In regard to explainability, the application of explainable machine learning methods to multi-stage hybrid architectures is yet to be developed because confidence-based routing results in complex decision-making that spans multiple machine learning algorithms, making it difficult to attribute the output to explainable input features. With regard to NIDS latency in production-ready deployments, especially in financial and industrial settings where low latency is crucial, more research is needed on the topic. Specifically, latency metrics should be provided in addition to accuracy scores in future studies. Lastly, NIDS must address the issues of non-IID data and model poisoning that federated learning algorithms face generally.

8. **CONCLUSION**

Indeed, machine learning has completely revolutionized the domain of network intrusion detection, allowing for the design of adaptive, intelligent solutions that learn from data to discover threats even when they differ substantially from any known attacks. In this review, we have reviewed the major algorithmic paradigms that have been used for NIDS development, such as traditional supervised classification methods, ensemble algorithms, neural networks and deep architectures, as well as more complex multi-stage pipelines that combine different approaches. We have reviewed the datasets typically employed to test and evaluate such approaches as well as some problems with this paradigm of assessment. The challenges of imbalanced class ratios, real-time latency, adaptability to encrypted traffic, and generalizability to new environments represent an agenda that goes far beyond improving the accuracy achieved by the existing systems on popular benchmarks. Federated learning, explainable AI, and graph-based detection present interesting new opportunities in this direction.

REFERENCES

1. Z. Ahmad, A. S. Khan, C. W. Shiang, J. Abdullah, and F. Ahmad (2020), "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," Trans. Emerg. Telecommun. Technol., vol. 32, no. 1, pp. 1–29.
2. F. Anis, M. Alabdullatif, S. Aljbli, and M. Hammoudeh (2025), "A survey on the applications of deep learning in network intrusion detection systems to enhance network security," IEEE Access, vol. 13, pp. 185357–185373.
3. A. Pinto, L.-C. Herrera, Y. Donoso, and J. A. Gutierrez (2023), "Survey on intrusion detection systems based on machine learning techniques for the protection of critical infrastructure," Sensors, vol. 23, no. 5, p. 2415.
4. Z. Dai, L. Y. Por, Y.-L. Chen, J. Yang, C. S. Ku, R. Alizadehsani, and P. Plawiak (2024), "An intrusion detection model to detect zero-day attacks in unseen data using machine learning," PLOS ONE, vol. 19, no. 9, p. e0308469.
5. T. H. Sow and M. Adda (2025), "Enhancing IDS performance through a comparative analysis of random forest, XGBoost, and deep neural networks," SSRN Preprint.
6. A. E. Onyebueke, A. A. David, and S. Munu (2023), "Network intrusion detection system using XGBoost and random forest algorithms," Asian J. Pure Appl. Math., vol. 5, no. 1, pp. 321–335.
7. Z. Xu et al. (2025), "Deep learning-based intrusion detection systems: A survey," arXiv:2504.07839.
8. M. Hozouri et al.(2025), "A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges," Discover Artif. Intell., vol. 5, p. 314.
9. H. Dinh, W. Zong, and Y.-W. Chow (2026), "Investigating oversampling techniques to mitigate class imbalance in network intrusion detection datasets," Pragmatic Cybersecurity, vol. 1.
10. Z. Zhou et al. (2023), "Hybrid CNN-LSTM for network intrusion detection," in Proc. IEEE Int. Conf. Big Data.
11. M. N. Alam et al.(2025), "A review of deep learning applications in intrusion detection systems: Overcoming challenges in spatiotemporal feature extraction and data imbalance," Appl. Sci., vol. 15, no. 3, p. 1552.
12. E. Tufan, C. Tezcan, and C. Acarturk (2021), "Anomaly-based intrusion detection by machine learning: A case study on probing attacks to an institutional network," IEEE Access, vol. 9, pp. 50078–50092.
13. Kikissagbe et al. (2024), "Improving DoS attack detection in IoT systems using SMOTE and ensemble classifiers," in Proc. IEEE Int. Conf. Cybersecurity,.
14. A. Hossain et al.(2024), "Machine learning based multi-stage intrusion detection using stacking ensemble on UNSW-NB15," in Proc. IEEE Int. Conf. Computing and Communication Technologies, pp. 88–94.
15. Z. Zoghi and G. Serpen (2024), "UNSW-NB15 computer security dataset: Analysis through visualization," Security and Privacy, vol. 7, p. e331.
16. B. Zoph et al. (2024), "Addressing class imbalance in network intrusion detection: A comprehensive evaluation of machine learning approaches," Electronics, vol. 14, no. 1, p. 69.
17. H. Wei et al. (2024), "HEN: Hybrid ensemble network with SHAP-guided intrusion detection using LightGBM and autoencoder-LSTM," Expert Syst. Appl.