

University of midocean

Faculty of Computer Science & Information Technology
Master of Science in Artificial Intelligence

Machine Learning-Based Intrusion Detection Systems
in Network Security:
A Comparative Literature Review

Assignment 2 — Problem Statement & Research Objectives

Course Code	AI800
Course Title	Research Methodology in Computer Science
Assignment	Assignment 2 — Problem Statement & Research Objectives
Lecturer	Dr. Mohammed Anbar
Student Name	Ibrahim Saber
Student ID	261002585
Trimester	First Trimester 2026
Submission Date	21 March 2026
Weighting	30%

1. Research Title

Machine Learning-Based Intrusion Detection Systems in Network Security: A Comparative Literature Review

2. Problem Statement

2.1 Problem Domain

The domain of this research is network security, specifically focusing on the design, evaluation, and deployment of Intrusion Detection Systems (IDS). An IDS is a critical component of modern cybersecurity infrastructure, responsible for monitoring network traffic and identifying malicious or anomalous behaviour in real time. With the exponential growth of internet-connected devices and the escalating sophistication of cyberattacks, securing digital networks has become a paramount concern for governments, enterprises, and academic institutions alike (Khraisat et al., 2019).

2.2 Description of the Problem

Traditional signature-based IDS rely on predefined patterns of known attacks and are therefore inherently unable to detect novel or zero-day threats. To overcome this limitation, researchers have increasingly turned to machine learning (ML) and deep learning (DL) techniques, which can learn complex, non-linear patterns from network traffic data without requiring explicit rule definitions. Despite considerable progress in this direction, two fundamental problems remain unsolved.

First, there is a persistent lack of realistic, standardised evaluation benchmarks. The most widely used dataset — KDD CUP 99 — contains approximately 78% duplicate records in its training set, a flaw that causes classifiers to achieve artificially inflated performance metrics and limits their real-world applicability (Tavallae et al., 2009). Although NSL-KDD was proposed as a remedy, it is also based on simulated traffic from a 1998 DARPA experiment and does not reflect contemporary attack vectors such as advanced persistent threats (APTs), encrypted malware, or Internet of Things (IoT) intrusions (Khraisat et al., 2019).

Second, existing ML/DL-based IDS models suffer from a fundamental tension between detection accuracy, computational efficiency, and model interpretability. Studies such as Vinayakumar et al. (2019) and Yin et al. (2017) demonstrate that deep neural architectures — including LSTM and CNN-LSTM hybrids — can achieve detection accuracies exceeding 99% on benchmark datasets. However, these models function as opaque black boxes and impose high computational costs, making them unsuitable for real-time deployment on resource-constrained network hardware. While Faker and Dogdu (2019) proposed a Big Data-integrated pipeline using Apache Spark to address the scalability challenge, their approach introduces significant infrastructure complexity and was validated on only a single dataset. Furthermore, none of the reviewed studies evaluated model robustness against adversarial attacks, where malicious actors deliberately craft network packets to evade ML-based detectors.

2.3 Significance of the Problem

The significance of this problem is underscored by the scale and consequences of network intrusions. Cybersecurity Ventures estimated that the global cost of cybercrime would exceed USD 8 trillion in 2023, with a projected rise to USD 10.5 trillion annually by 2025. Network intrusions — including Distributed Denial of Service (DDoS) attacks, ransomware, and advanced persistent threats — represent a substantial portion of these losses. An IDS that is either inaccurate, too slow, or easily evaded by adversarial manipulation provides only an illusion of security, exposing organisations to catastrophic data breaches, service outages, and regulatory penalties.

Moreover, the rapid expansion of IoT ecosystems, cloud computing environments, and encrypted network protocols has dramatically enlarged the attack surface that IDS must monitor. As highlighted by Khraisat et al. (2019), the existing body of IDS research has largely neglected these emerging network environments, rendering most proposed solutions inadequate for the challenges of modern network security.

2.4 Existing Solutions and Their Limitations

The literature reveals a clear progression of IDS methodologies, each with notable limitations:

- Signature-based IDS (e.g., Snort, Suricata) offer high precision for known attacks but fail entirely against novel or polymorphic threats, as they require constant manual signature updates.
- Statistical ML-based IDS (e.g., SVM, Decision Trees, Random Forests) improve generalisation over signature methods but rely heavily on handcrafted feature engineering and struggle with high-dimensional, time-sequential traffic data.
- Deep learning IDS (Yin et al., 2017; Shone et al., 2018; Vinayakumar et al., 2019) eliminate the need for manual feature engineering and achieve high accuracy on benchmarks. However, they lack interpretability, incur high computational costs, and have not been tested against adversarial perturbations.
- Scalable DL-IDS (Faker & Dogdu, 2019) address throughput limitations through distributed big-data frameworks, but introduce deployment complexity and do not resolve the interpretability or adversarial robustness gaps.

Across all categories, a critical common weakness is the exclusive use of outdated or synthetic benchmark datasets for evaluation, which prevents conclusions about real-world performance from being drawn with confidence (Tavallaei et al., 2009; Khraisat et al., 2019).

2.5 Justification for Further Research

The gaps identified above — unrealistic evaluation benchmarks, poor interpretability, high computational demands, and the absence of adversarial robustness assessments — collectively constitute a significant barrier to the adoption of ML/DL-based IDS in real-world production environments. There is, therefore, a compelling and urgent need for research that:

1. Proposes and validates an ML-based IDS framework evaluated against both established benchmarks (NSL-KDD, UNSW-NB15) and contemporary, real-world traffic captures to ensure ecological validity.
2. Designs a detection architecture that achieves a balance between high detection accuracy and computational efficiency, enabling practical deployment in resource-constrained network environments.
3. Incorporates interpretability mechanisms — such as Explainable AI (XAI) techniques — to provide transparent, auditable justification for intrusion classifications, satisfying the requirements of security practitioners and regulatory frameworks.
4. Evaluates the adversarial robustness of the proposed model against state-of-the-art evasion attacks to measure its resilience under realistic threat conditions.

This research aims to address the above gaps, thereby contributing a more robust, interpretable, and practically deployable ML-based IDS framework to the field of network security.

3. Research Objectives

The primary goal of this research is to propose and evaluate an interpretable, efficient, and adversarially robust machine learning-based Intrusion Detection System (ML-IDS) capable of detecting both known and novel network intrusions with high accuracy across diverse and realistic network environments.

To achieve this primary goal, the following specific research objectives are defined:

Objective 1: To analyse and critically compare the methodological approaches, evaluation datasets, and performance metrics employed in existing ML/DL-based IDS research, in order to identify the most significant technical gaps and limitations that prevent current solutions from being deployed in real-world network environments.

This objective directly addresses the problem of fragmented and inconsistent evaluation methodologies identified in the literature review, particularly the over-reliance on outdated, synthetic datasets such as KDD CUP 99 (Tavallae et al., 2009) and the absence of standardised cross-study comparisons (Khraisat et al., 2019). Measurability: Success will be assessed by the production of a structured comparative taxonomy encompassing at least fifteen peer-reviewed IDS studies, documenting datasets used, accuracy metrics, and identified limitations.

Objective 2: To design a hybrid ML-based IDS architecture that integrates a feature extraction module (convolutional layers for spatial pattern recognition) with a temporal modelling module (LSTM layers for sequential traffic analysis), optimised for both detection accuracy and computational efficiency on high-speed network traffic.

This objective targets the accuracy–efficiency tension highlighted by Vinayakumar et al. (2019) and Faker and Dogdu (2019). The CNN-LSTM hybrid design is motivated by the superior performance of such architectures in prior benchmarks. Measurability: The architecture will be evaluated by its classification accuracy (F1-score, precision, recall) and inference latency (milliseconds per sample), targeting an F1-score above 0.95 and real-time processing capability (< 10 ms per packet on standard server hardware).

Objective 3: To develop a dataset augmentation and preprocessing pipeline that combines publicly available benchmark datasets (NSL-KDD, UNSW-NB15, CICIDS2017) with curated real-world traffic captures, including samples of encrypted traffic and IoT device communication patterns, to produce a more representative training and evaluation corpus.

This objective directly responds to the foundational dataset problem articulated by Tavallaei et al. (2009) and reinforced by Khraisat et al. (2019). By constructing a multi-source dataset, the research ensures that model performance generalises beyond legacy benchmarks. Measurability: Success will be quantified by the number of unique attack categories represented, the proportion of encrypted traffic samples included, and the statistical distribution characteristics of the augmented dataset relative to real network traffic captures.

Objective 4: To integrate Explainable Artificial Intelligence (XAI) techniques — specifically SHAP (SHapley Additive exPlanations) — into the proposed IDS to provide feature-level explanations for each intrusion classification, thereby improving transparency and supporting security analyst decision-making.

This objective addresses the interpretability gap identified across all experimental papers reviewed (Yin et al., 2017; Shone et al., 2018; Vinayakumar et al., 2019; Faker & Dogdu, 2019). SHAP values provide mathematically grounded, model-agnostic explanations consistent with the requirements of security audits and compliance frameworks. Measurability: The quality of explanations will be evaluated through a user study with network security professionals, measuring explanation fidelity, actionability, and user trust on standardised scales.

Objective 5: To evaluate the adversarial robustness of the proposed ML-IDS against representative evasion attack strategies — including FGSM (Fast Gradient Sign Method) and PGD (Projected Gradient Descent) perturbations — and to propose adversarial training countermeasures that maintain detection performance under adversarial conditions.

None of the six reviewed papers assessed adversarial robustness, constituting a critical gap given that sophisticated adversaries actively attempt to evade ML-based detectors. This objective ensures the proposed system is stress-tested under realistic threat conditions. Measurability: Robustness will be quantified by comparing F1-scores before and after adversarial attack exposure, targeting a performance degradation of less than 5% under standardised FGSM and PGD perturbation budgets.

4. Literature Review

The following six papers were selected because they represent seminal or highly cited contributions to the field of ML-based network intrusion detection. The selection spans from 2009 to 2019 to capture the progression of methodological sophistication, from foundational dataset analysis to advanced deep learning architectures deployed on big-data platforms.

4.1 A Detailed Analysis of the KDD CUP 99 Data Set

Reference:

Tavallaei, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A Detailed Analysis of the KDD CUP 99 Data Set. Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), pp. 1–6.

Summary:

This paper provides a rigorous statistical examination of the KDD CUP 99 dataset, which was — and remains — one of the most widely used benchmarks for evaluating intrusion detection systems. The authors discovered that approximately 78% of the records in the training set are duplicates and around 75% of the test set records are duplicated, severely biasing evaluation metrics. As a remedy, they proposed and released the NSL-KDD dataset, which eliminates redundant records, balances difficulty levels across attack categories, and provides a more reliable benchmark for future IDS research. The methodology involved statistical comparison of class distributions, frequency analysis of duplicate records, and evaluation of several classifiers (C4.5, NaiveBayes, and NBTree) on both KDD CUP 99 and NSL-KDD to demonstrate improved generalisation on the cleaner dataset.

Critical Analysis:

Strengths: The paper addresses a fundamental flaw that undermined years of prior IDS research, making it an indispensable methodological contribution. The release of NSL-KDD as a public resource added significant practical value to the research community. The statistical evidence presented is transparent and reproducible. **Weaknesses:** The paper does not introduce a new intrusion detection algorithm; it is purely a dataset critique. Moreover, both datasets were constructed from simulated network traffic, which limits ecological validity. The dataset is also outdated and does not reflect modern attack vectors such as APTs or IoT intrusions.

4.2 A Deep Learning Approach for Intrusion Detection Using RNNs

Reference:

Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks. IEEE Access, 5, 21954–21961.

Summary:

Yin et al. proposed an intrusion detection model based on Recurrent Neural Networks (RNN), specifically leveraging the Long Short-Term Memory (LSTM) architecture to capture temporal dependencies in network traffic sequences. Trained and evaluated on NSL-KDD, the RNN-IDS outperformed SVM and Random Forests, achieving binary classification accuracy of up to 99.61% and multi-class accuracy of 83.28%. The study also provided a sensitivity analysis of hyper-parameters such as the number of hidden units and learning rate.

Critical Analysis:

Strengths: The application of LSTM to model temporal patterns in network traffic is a meaningful innovation over stateless classifiers. Thorough hyper-parameter analysis enhances reproducibility. **Weaknesses:** Evaluation is entirely confined to NSL-KDD, raising questions about external validity. The computational overhead of LSTM models is acknowledged but not rigorously quantified in terms of real-time feasibility. Multi-class accuracy of 83.28% for rare attack categories (R2L, U2R) may be insufficient for production deployment.

4.3 A Deep Learning Approach to Network Intrusion Detection

Reference:

Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A Deep Learning Approach to Network Intrusion Detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(1), 41–50.

Summary:

Shone et al. introduced a non-symmetric deep autoencoder (NDAE) combined with a Random Forest classifier. The NDAE performs unsupervised feature learning without requiring decoding, reducing computational burden. Experiments on KDD CUP 99 and NSL-KDD demonstrated 97.85% accuracy on KDD CUP 99 and 85.42% on NSL-KDD, with training times reduced by approximately 15% compared to stacked autoencoder approaches.

Critical Analysis:

Strengths: The NDAE design is computationally efficient. The hybrid unsupervised-supervised approach reduces dependence on labelled data. **Weaknesses:** Exclusive reliance on benchmark datasets limits generalisability. The notably lower performance on NSL-KDD (85.42%) compared to KDD CUP 99 (97.85%) suggests potential over-fitting to the older dataset. Encrypted network traffic is not considered.

4.4 Deep Learning Approach for Intelligent Intrusion Detection System

Reference:

Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep Learning Approach for Intelligent Intrusion Detection System. *IEEE Access*, 7, 41525–41550.

Summary:

This study systematically evaluates multiple DNN architectures — MLP, CNN, RNN/LSTM, and CNN-LSTM hybrids — on KDD CUP 99 and UNSW-NB15. The CNN-LSTM hybrid achieved the highest accuracy: 99.44% on KDD CUP 99 and 88.79% on UNSW-NB15. The paper also examined the effect of depth, dropout regularisation, and batch normalisation on detection performance.

Critical Analysis:

Strengths: The breadth of comparing seven DNN architectures across two datasets provides an exceptionally valuable empirical reference. Inclusion of UNSW-NB15 enhances practical relevance. **Weaknesses:** Deep networks function as black boxes with no interpretability. High computational cost makes real-time deployment on constrained hardware challenging. Adversarial robustness is not evaluated.

4.5 Survey of Intrusion Detection Systems: Techniques, Datasets and Challenges

Reference:

Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of Intrusion Detection Systems: Techniques, Datasets and Challenges. *Cybersecurity*, 2(1), 20.

Summary:

This systematic literature review of 124 IDS papers (2000–2019) established a taxonomy categorising IDS into signature-based, anomaly-based, and hybrid approaches. The authors compared benchmark datasets and identified persistent challenges: lack of standardised evaluation metrics, absence of real-world validated datasets, limited focus on IoT-specific intrusions, and poor handling of zero-day attacks.

Critical Analysis:

Strengths: The comprehensive scope and structured taxonomy make it an authoritative reference. Identification of the dataset standardisation problem is particularly valuable. Methodologically rigorous with clearly defined inclusion/exclusion criteria. **Weaknesses:** No original experimental data. Findings may become outdated rapidly. The review does not deeply examine privacy implications or adversarial robustness of reviewed methods.

4.6 Intrusion Detection Using Big Data and Deep Learning Techniques

Reference:

Faker, O., & Dogdu, E. (2019). Intrusion Detection Using Big Data and Deep Learning Techniques. Proceedings of the 2019 ACM Southeast Conference (ACM SE '19), pp. 86–93.

Summary:

Faker and Dogdu proposed an IDS pipeline integrating Apache Spark with CNN and RNN/LSTM models, evaluated on UNSW-NB15. The distributed architecture reduced processing time for high-volume traffic streams. CNN achieved 98.36% accuracy and RNN achieved 97.13%, demonstrating the feasibility of scalable, production-oriented DL-IDS deployment.

Critical Analysis:

Strengths: Addressing scalability is a critical practical contribution. The use of UNSW-NB15 lends more credibility than KDD-only studies. The Spark integration demonstrates a clear pathway to production deployment. **Weaknesses:** Confined to a single dataset. Infrastructure complexity creates deployment barriers. Adversarial robustness and model explainability are not addressed.

4.7 Connections and Conflicts Between Papers

Convergent Findings:

- **Dataset dependency:** Five of the six papers use KDD CUP 99 or NSL-KDD, reinforcing the centrality of these benchmarks and — as Tavallae et al. (2009) cautioned — the risk of inflated performance metrics due to dataset flaws.
- **Deep learning superiority:** Yin et al. (2017), Shone et al. (2018), Vinayakumar et al. (2019), and Faker & Dogdu (2019) all demonstrate that DL architectures outperform traditional ML classifiers, establishing a clear research consensus.
- **Benchmark inadequacy:** Khraisat et al. (2019) and Tavallae et al. (2009) both emphasise that available benchmarks do not adequately represent real-world network conditions.

Conflicting Findings and Tensions:

- **Performance inconsistency across datasets:** Shone et al. report 97.85% on KDD CUP 99 but only 85.42% on NSL-KDD; Vinayakumar et al. achieve 99.44% on KDD but 88.79% on UNSW-NB15. These discrepancies highlight dataset-dependent performance and challenge accuracy as a universal measure.
- **Scalability vs. accuracy trade-off:** Faker & Dogdu prioritise scalability through big-data integration, while other studies focus on maximising accuracy without considering deployment at scale, reflecting a fundamental tension between optimisation and operational practicality.

5. Summary Table of Reviewed Papers

Year	Author(s)	Methodology	Key Findings	Weaknesses / Limitations	Citation
2009	Tavallae et al.	Statistical analysis & pre-processing of KDD CUP 99 dataset	Identified severe redundancy in training/test sets; proposed NSL-KDD as a cleaner benchmark	No new ML model proposed; dataset constructed from simulated — not real-world — traffic	Tavallae et al. (2009)
2017	Yin et al.	RNN / LSTM on NSL-KDD dataset; binary & multi-class classification	RNN-IDS outperformed SVM and Random Forest; binary accuracy up to 99.61%	Evaluated only on NSL-KDD; no real-time validation; high computational cost	Yin et al. (2017)
2018	Shone et al.	Non-symmetric Deep Auto-encoder (NDAE) + Random Forest hybrid	97.85% accuracy on KDD CUP 99; ~15% faster training than stacked autoencoders	Results not generalised beyond benchmark datasets; encrypted traffic not addressed	Shone et al. (2018)
2019	Vinayakumar et al.	Systematic comparison of MLP, CNN, RNN/LSTM, CNN-LSTM on KDD CUP 99 & UNSW-NB15	CNN-LSTM hybrid achieved 99.44% (KDD) and 88.79% (UNSW-NB15); best overall accuracy	Black-box DNN; high resource requirements; adversarial robustness not evaluated	Vinayakumar et al. (2019)
2019	Khraisat et al.	Systematic literature review of 124 IDS papers; taxonomy of techniques & datasets	Comprehensive taxonomy of anomaly, signature & hybrid IDS; identified gaps in IoT and zero-day detection	No original experimental data; survey-based findings may become outdated rapidly	Khraisat et al. (2019)
2019	Faker & Dogdu	Apache Spark (Big Data) + CNN/RNN on UNSW-NB15	CNN accuracy of 98.36%; demonstrated scalable, production-oriented DL-IDS pipeline	Single dataset; Spark-DL infrastructure complexity barriers for small organisations	Faker & Dogdu (2019)

6. Synthesis of Information

The six reviewed papers collectively construct a coherent narrative of progress, persistent challenges, and emerging research directions in ML-based IDS. The field's trajectory moves from foundational dataset critique (Tavallaei et al., 2009) through empirical validation of deep learning superiority (Yin et al., 2017; Shone et al., 2018), to comprehensive cross-architecture evaluation (Vinayakumar et al., 2019), a synthesising survey identifying persistent gaps (Khraisat et al., 2019), and a scalable production framework (Faker & Dogdu, 2019).

Three cross-cutting themes emerge. First, the dataset challenge remains unresolved: no universally accepted benchmark capturing contemporary attack patterns exists, inhibiting meaningful cross-study comparisons. Second, there is a recurring tension between model complexity and operational practicality. Third, interpretability and adversarial robustness represent critical blind spots: none of the experimental studies assess performance under adversarial attacks, and the black-box nature of deep networks conflicts with security audit requirements.

These three themes directly motivate the five research objectives defined in Section 3. The proposed research aims to address all identified gaps simultaneously, thereby contributing a more robust, interpretable, and practically deployable ML-based IDS framework to the field of network security.

7. References

- Faker, O., & Dogdu, E. (2019). Intrusion detection using big data and deep learning techniques. *Proceedings of the 2019 ACM Southeast Conference*, pp. 86–93. <https://doi.org/10.1145/3299815.3314439>
- Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity*, 2(1), 20. <https://doi.org/10.1186/s42400-019-0038-7>
- Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A deep learning approach to network intrusion detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(1), 41–50. <https://doi.org/10.1109/TETCI.2017.2772792>
- Tavallae, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. *Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, pp. 1–6. <https://doi.org/10.1109/CISDA.2009.5356528>
- Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. *IEEE Access*, 7, 41525–41550. <https://doi.org/10.1109/ACCESS.2019.2895334>
- Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access*, 5, 21954–21961. <https://doi.org/10.1109/ACCESS.2017.2762418>