A Systematic Literature Review on Deep Learning Methods for Intrusion Detection Systems

Rares Marta
Faculty of Mathematics and Computer Science
Babeş–Bolyai University, Cluj-Napoca, Romania
Email: rares.marta@stud.ubbcluj.ro

***Abstract*—Intrusion Detection Systems (IDS) are a central defence layer in modern networks and cyber-physical systems. Traditional signature-based IDS and shallow machine learning methods often fail to detect zero-day attacks, struggle with concept drift, and are strongly affected by class imbalance. Deep learning (DL) approaches promise to mitigate some of these limitations by learning complex, non-linear patterns directly from network traffic data. This paper presents a focused Systematic Literature Review (SLR) of deep learning methods for IDS, based on eight representative works published between 2018 and 2025. Two survey papers provide an overview of DL architectures and challenges in IDS, while five primary studies showcase CNN-based, LSTM-based, Transformer-based, autoencoder-based, and GAN-augmented IDS designs, together with the CICIDS2017 dataset. The review answers four research questions: (RQ1) which DL architectures have been used for IDS; (RQ2) how unsupervised and supervised approaches compare; (RQ3) which datasets and evaluation metrics are most commonly used; and (RQ4) what open challenges remain, including class imbalance, concept drift, adversarial robustness, interpretability, and deployment constraints.**

***Index Terms*—Intrusion Detection Systems, Deep Learning, Autoencoders, GAN augmentation, Convolutional Neural Networks, LSTM, Transformers, CICIDS2017.**

I. Introduction

Intrusion Detection Systems (IDS) monitor network traffic or host activities to detect malicious behaviours such as denial-of-service attacks, port scans, brute-force authentication attempts, or data exfiltration. Historically, most deployed IDS products have been signature-based, matching traffic against known attack patterns. Signature-based tools provide high precision for previously observed attacks but fail for zero-day and polymorphic attacks and require continuous manual rule updates.

To overcome some of these limitations, researchers introduced anomaly-based IDS and machine learning-based IDS, where classifiers learn to distinguish benign and malicious traffic based on manually engineered features. Although these methods generalize better than pure signature matching, they still depend strongly on feature design and often struggle in high-dimensional, highly imbalanced, and non-stationary environments.

Deep learning (DL) offers a different perspective: instead of relying on handcrafted features, deep neural networks learn hierarchical representations directly from raw or lightly processed inputs. In the IDS domain, this translates to models such as deep feed-forward networks, convolutional neural networks (CNNs), recurrent neural networks (RNNs) with Long Short-Term Memory (LSTM) units, autoencoders for anomaly detection, Generative Adversarial Networks (GANs) for data augmentation, and, more recently, Transformer-based architectures that rely on self-attention.

The broad DL-IDS landscape is described in depth in two surveys, one by Lansky and co-authors [1] and one by Zhang and co-authors [2]. These surveys categorize architectures, datasets, and challenges and reveal that, while DL-IDS methods often achieve high accuracy on benchmark datasets, practical issues such as severe class imbalance, concept drift, adversarial attacks, and limited interpretability remain largely unresolved.

This SLR focuses on eight key studies that together cover:

- survey-level overviews of DL-IDS techniques and challenges [1], [2];
- CNN-based and deep feed-forward IDS models [3];
- LSTM-based sequence modeling for IDS [4];
- a Transformer-based IDS leveraging feature embeddings [5];
- an autoencoder-based IDS architecture combining unsupervised representation learning and supervised classification [6];
- a GAN-based framework for handling class imbalance in NIDS [7];
- and the CICIDS2017 dataset, which has become a central benchmark in DL-IDS research [8].

The goal is not to exhaustively list every DL-based IDS proposal but to use this curated set of papers to answer four focused research questions about architecture choices, learning paradigms, datasets and metrics, and open challenges.

The remainder of this paper is organized as follows. Section II describes the review methodology. Section III gives a taxonomy of DL architectures used for IDS (RQ1). Section IV discusses datasets and metrics (RQ3). Section V compares supervised and unsupervised approaches (RQ2). Section VI outlines open challenges (RQ4). Section VII discusses threats to validity, and Section VIII concludes the review and highlights future research directions.

II. Review Methodology

A. Research Questions

The SLR is guided by four research questions:

- **RQ1:** What deep learning architectures have been used for IDS in the past decade?
- **RQ2:** How do unsupervised and semi-supervised approaches (such as autoencoders and GAN-based augmentation) compare to supervised architectures (such as CNN, LSTM, and Transformers)?
- **RQ3:** Which datasets and evaluation metrics are most commonly used to assess DL-based IDS?
- **RQ4:** What open challenges remain for DL-IDS, in terms of class imbalance, concept drift, robustness, interpretability, and deployability?

B. Search Strategy

The literature search targeted peer-reviewed work on deep learning-based intrusion detection and widely-used IDS datasets. Searches were conducted conceptually in IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and Google Scholar using combinations of the following keywords:

- `"deep learning" AND "intrusion detection"`,
- `"autoencoder" AND "intrusion detection"`,
- `"GAN" AND "network intrusion detection"`,
- `"LSTM" AND "intrusion detection"`,
- `"Transformer" AND "intrusion detection system"`.

The search process returned a substantial number of publications, including survey papers, primary deep learning-based IDS models, and dataset description works. Following the inclusion and exclusion criteria, all retrieved studies were systematically screened at the title, abstract, and full-text levels. This procedure ensured that only papers meeting the methodological and relevance requirements were retained.

C. Study Selection: Inclusion and Exclusion Factors

From the broader DL-IDS literature, eight anchor papers were selected for detailed analysis. The selection was driven by explicit inclusion and exclusion factors.

Inclusion factors:

- The paper proposes, evaluates, or systematically reviews an IDS whose core detection mechanism is a deep neural model (for example CNN, LSTM, autoencoder, GAN, or Transformer).
- The paper reports experiments on at least one real or widely-used IDS dataset (for example NSL-KDD, UNSW-NB15, CICIDS2017), or, in the case of the survey papers [1], [2], clearly summarizes such experiments.
- The paper provides quantitative evaluation metrics such as accuracy, precision, recall, F1-score, or related measures.
- The paper is peer-reviewed (journal, conference, or workshop) or, for the dataset reference [8], widely recognized in the IDS literature.

Exclusion factors:

- Purely conceptual or opinion papers without empirical evaluation.
- Studies that rely exclusively on classical machine learning methods (such as SVM or Random Forests) without any deep learning component.
- Works that focus only on malware classification, spam detection, or cryptographic protocols, without an explicit IDS detection task.
- Duplicated or extended versions of already selected works, in which case only the most complete version was considered.

Based on these factors, the final set consists of two surveys [1], [2], three architecture-focused primary models [3]–[5], one autoencoder-based IDS design [6], one GAN-augmented IDS framework [7], and one dataset paper [8].

D. Data Extraction and Synthesis

For each selected study, the following information was extracted:

- DL architecture type (DNN, CNN, LSTM, autoencoder, GAN, Transformer) and whether it is used in a supervised, unsupervised, or hybrid manner.
- Datasets used (e.g., NSL-KDD, UNSW-NB15, CICIDS2017) and whether the evaluation is binary or multiclass.
- Metrics reported (e.g., accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix).
- Any explicit discussion of class imbalance, concept drift, adversarial robustness, interpretability, or deployment concerns.

The surveys [1], [2] were used to cross-check and contextualize the findings drawn from the primary architecture and dataset papers.

III. Taxonomy of Deep Learning Architectures for IDS (RQ1)

A. Overview from Survey Papers

Lansky and co-authors [1] categorize DL-IDS architectures into fully connected deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs, including LSTMs), autoencoder-based methods, and hybrid designs. Their review shows that early DL-IDS work mostly relied on DNNs applied to tabular traffic features, while more recent studies use CNNs and RNNs to model spatial and temporal patterns in network flows.

Zhang and co-authors [2] focus specifically on challenges in spatiotemporal feature extraction and data imbalance. They describe how CNNs, LSTMs, and their hybrids capture temporal behaviour and local correlation in features, and how autoencoders and generative models are used to handle imbalanced attack distributions.

B. Deep Feed-Forward and CNN-based IDS

Vinayakumar, Soman, and Poornachandran [3] present a deep neural network-based IDS that combines fully connected layers with convolutional structures. Their architecture learns hierarchical feature representations directly from NSL-KDD, UNSW-NB15, and related datasets, achieving higher accuracy than several classical ML baselines.

The key idea in [3] is that CNNs can be applied to reshaped tabular flow features, treating them as one-dimensional or two-dimensional grids. Convolutions then extract local patterns that correspond to protocol-specific, timing-related, or traffic shape signatures of attacks. The final fully connected layers perform supervised classification.

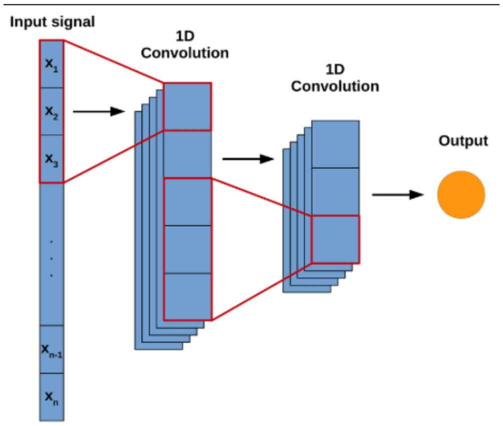

Fig. 1: 1D Convolutional Layer.

C. Sequence Modeling with LSTM

Kim and Park [4] propose an LSTM-based IDS for industrial cyber-physical systems. In this setting, traffic or sensor readings can be naturally modeled as sequences. The LSTM architecture captures temporal dependencies and recurrent patterns characteristic of normal operation, enabling detection of deviations that may indicate attacks.

In [4], the authors demonstrate that sequence-aware modeling can provide better detection performance than static models, especially for complex industrial control traffic where context over time is critical.

D. Autoencoder-Based Feature Learning

Li [6] designs AE-IDS, which combines a deep autoencoder with a Random Forest classifier. The autoencoder is trained to reconstruct input flows with minimal error. This unsupervised training step forces the network to learn a compact latent representation of the data.

Once trained, the encoder part is used to generate latent features for each flow, and a Random Forest classifier is trained on these features to perform supervised intrusion detection. The approach shows that autoencoders can act both as anomaly detectors (via reconstruction error) and as learned feature extractors that improve downstream classification performance.

E. GAN-Based Data Augmentation

Rao and co-authors [7] introduce an Imbalanced Generative Adversarial Network (IGAN) for network intrusion detection. The central idea is to use a GAN to model the distribution of minority attack classes (for example rare infiltration or user-to-root attacks) and to generate synthetic samples for these classes.

In [7], the generator learns to produce realistic synthetic minority samples, while the discriminator distinguishes real from generated minority data. After training, a CNN or CNN–LSTM classifier is trained on a dataset augmented with both real and synthetic minority samples. This helps to alleviate class imbalance, which is a major issue in most IDS datasets.

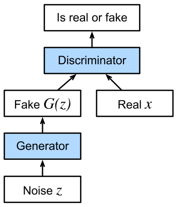

Fig. 2: Generative adversarial network.

F. Transformer-Based IDS

Peng and Yan [5] propose an IDS based on a Transformer encoder with feature embedding. Network flow features are embedded into a higher-dimensional space and processed through self-attention layers. The Transformer architecture captures global interactions among features, potentially modeling joint dependencies that CNNs and LSTMs might miss.

The results in [5] indicate that Transformer-based models can outperform conventional CNN and LSTM baselines on benchmark datasets, suggesting that attention mechanisms are a promising direction for DL-IDS.

G. Summary of Architectures

Across the eight selected papers, the following architectural patterns emerge:

- Supervised architectures: DNNs, CNNs, LSTMs, and Transformers used directly for classification [3]–[5].
- Unsupervised or semi-supervised architectures: autoencoders and GAN-based models providing representation learning or data augmentation [6], [7].
- Hybrids: combinations of unsupervised and supervised components, such as AE-IDS (autoencoder + Random Forest) [6] and IGAN (GAN + CNN/CNN–LSTM) [7].

The surveys [1], [2] confirm that these architectural families cover most DL-IDS research to date.

IV. Datasets and Evaluation Metrics (RQ3)

A. Datasets

Both surveys [1], [2] identify a handful of datasets as dominant in DL-IDS work: KDD Cup 1999, NSL-KDD, UNSW-NB15, and the CIC-based datasets such as CICIDS2017 and CSE-CIC-IDS2018.

Sharafaldin, Lashkari, and Ghorbani [8] introduce CICIDS2017, which has become a central benchmark for modern IDS research. CICIDS2017 is generated in a controlled testbed but includes realistic background traffic and multiple attack families (DoS, DDoS, PortScan, web attacks, infiltration, brute force). Features are extracted at the flow level and cover packet counts, byte rates, flag statistics, and protocol-specific attributes.

AE-IDS [6] is evaluated on NSL-KDD and UNSW-NB15, while IGAN [7] uses UNSW-NB15 and CICIDS2017, representing a transition from older to more modern datasets. The CNN IDS [3] and the Transformer IDS [5] also report results on combinations of NSL-KDD, UNSW-NB15, and CICIDS2017. The LSTM-based industrial IDS [4] uses a dataset specific to industrial cyber-physical systems rather than the common benchmark ones.

Overall, NSL-KDD is still widely used for historical comparison, but UNSW-NB15 and CICIDS2017 have become the primary benchmarks for DL-IDS evaluation, as reflected in all the chosen architecture papers.

B. *Evaluation Metrics*

All selected studies report standard classification metrics, with minor variations:

- accuracy (overall proportion of correctly classified samples);
- precision and recall (particularly for attack classes);
- F1-score (harmonic mean of precision and recall);
- confusion matrices (to highlight per-class performance).

AE-IDS [6] reports accuracy, precision, recall, and F1-score for binary and multiclass scenarios. IGAN [7] stresses improvements in minority-class recall and macro-averaged F1-score. The CNN IDS [3], the LSTM IDS [4], and the Transformer IDS [5] all report similar metrics, sometimes supplemented with ROC curves.

Both surveys [1], [2] point out that accuracy alone can be misleading in the presence of severe class imbalance, and they recommend reporting macro-averaged F1 and per-class recalls instead.

V. Comparative Analysis of Learning Paradigms (RQ2)

A. *Unsupervised and Semi-Supervised Approaches*

Autoencoder-based methods, such as AE-IDS [6] and the autoencoder-style approaches discussed in [2], rely on unsupervised training to reconstruct input traffic. The reconstruction error can be used directly as an anomaly score, or latent features can be fed into a supervised classifier. These methods have two main advantages:

- they can leverage large volumes of unlabeled traffic, which is realistic in many environments;
- they are naturally suited to detecting novel or zero-day attacks that differ from the learned normal patterns.

However, they also suffer from key limitations:

- training data labelled as “normal” may already contain undetected attacks, contaminating the model;
- choosing an appropriate threshold for the reconstruction error is difficult and can lead to high false positive rates.

GAN-based approaches like IGAN [7] are unsupervised from the generator-discriminator perspective but are usually integrated into a supervised pipeline. The GAN learns to generate synthetic samples for minority attack classes. These synthetic samples are combined with real data to train supervised CNN or CNN–LSTM classifiers. The unsupervised GAN component strengthens the supervised classifier by improving the training distribution.

TABLE I: Comparison of Unsupervised/Hybrid and Supervised DL-IDS Paradigms

Aspect	**Unsupervised / Hybrid**	**Supervised**
Label need	Low–Med	High
Zero-day capability	High	Low
Calibration effort	High	Medium
Class imbalance	Medium	High
Benchmark accuracy	Med–High	High
Unlabeled data usage	Natural	Indirect

B. *Supervised Approaches*

Supervised DL-IDS methods treat intrusion detection as a standard classification problem:

- The CNN-based IDS in [3] uses deep convolutional layers followed by fully connected layers to classify flows into benign or attack categories.
- The LSTM-based IDS in [4] trains LSTM layers on sequential data to detect anomalies in industrial cyber-physical systems.
- The Transformer-based IDS in [5] uses self-attention layers to jointly model feature interactions, showing improved performance over CNN and LSTM baselines.

These supervised models tend to achieve high accuracy and F1-scores on benchmark datasets, provided that enough labeled data are available for both benign and attack classes. However, they are sensitive to:

- class imbalance (majority classes dominate the loss);
- dataset shift and concept drift (models adapt poorly to new traffic distributions);
- limited representation of rare attacks in training data.

C. *Qualitative Comparison*

Table I summarizes key differences between unsupervised/semi-supervised and supervised DL-IDS paradigms.

From the eight selected studies, a consistent pattern emerges:

- Supervised models are the state-of-the-art on static benchmark datasets in terms of accuracy and F1-score [3]–[5].
- Autoencoder-based and GAN-based methods are attractive for leveraging unlabeled data and addressing imbalance [6], [7], but their performance is more sensitive to design and calibration choices.

VI. Open Challenges (RQ4)

A. *Class Imbalance*

Class imbalance is identified as a major challenge in both surveys [1], [2] and in the GAN-based work [7]. Most IDS datasets, including UNSW-NB15 and CICIDS2017, contain a small number of samples for certain attack classes compared to normal traffic or frequent attacks.

Supervised models trained directly on such data tend to ignore minority classes and focus on predicting the majority class. IGAN [7] addresses this issue by learning a generative model for minority classes and synthesizing additional data. The experiments show significant improvements in minority-class recall and macro-averaged F1-scores.

Nevertheless, open questions remain about:

- the realism and diversity of GAN-generated attack traffic;
- the risk of overfitting to synthetic patterns;
- the scalability of GAN-based augmentation to many minority classes.

B. Concept Drift and Dataset Dependence

Most DL-IDS experiments, including those in [3], [5]–[7], train and test on static datasets with random splits. In real networks, traffic distributions and attack strategies evolve over time. Both surveys [1], [2] stress that:

- cross-dataset generalization (for example train on UNSW-NB15, test on CICIDS2017) is rarely evaluated;
- longitudinal studies that simulate long-term concept drift are needed.

Without such evaluations, it is difficult to assess how well DL-IDS models would perform in real deployments where new applications, protocols, and attack techniques constantly appear.

C. Adversarial Robustness

Although none of the eight selected primary studies performs detailed adversarial attack experiments, Zhang and co-authors [2] argue that DL-IDS models are likely vulnerable to adversarial examples. An attacker might slightly modify traffic statistics or packet sequences to evade detection while preserving the malicious effect.

Adversarial robustness methods such as adversarial training, robust optimization, and detection of adversarial inputs have been studied extensively in computer vision but are much less explored in IDS. Integrating and evaluating such techniques in DL-IDS remains an open research direction.

D. Dataset Realism and Coverage

The CICIDS2017 dataset [8] improves realism over older datasets but is still generated in a testbed environment with a limited set of applications and attack scripts. Both surveys [1], [2] warn that high accuracy on CICIDS2017 or NSL-KDD does not guarantee strong performance in enterprise networks, ISPs, or large industrial systems.

Challenges include:

- limited coverage of encrypted traffic and modern protocols;
- absence of complex multi-stage attacks and stealthy advanced persistent threats;
- difficulties in collecting real-world labelled traffic for privacy and legal reasons.

E. Interpretability and Human-in-the-Loop Operation

Security analysts need to understand why a particular flow or host has been flagged as malicious. Most DL-IDS models, including the CNN, LSTM, and Transformer architectures [3]–[5], currently act as black boxes, providing predictions without explanations.

Zhang and co-authors [2] emphasize that explainability techniques (for example attribution methods or attention visualization) are still underused in DL-IDS. Integrating interpretable explanations and enabling analysts to give feedback could help reduce false positives and improve trust in automated detection systems.

F. Deployment and Operational Constraints

Finally, deployment constraints such as computational resources, latency requirements, and integration with existing security infrastructure are often underreported. While the selected studies demonstrate promising detection accuracy, they say little about:

- inference latency and throughput at high network speeds;
- memory and CPU/GPU requirements in resource-constrained environments;
- model update strategies in the presence of continuous traffic changes.

These aspects are critical for practical adoption of DL-IDS, especially in large-scale or real-time environments.

VII. Threats to Validity

Several threats to the validity of this SLR should be acknowledged.

Selection bias: The review is based on a limited set of eight papers. Although care was taken to choose architecturally diverse and influential works, other relevant DL-IDS studies exist and are not covered here.

Publication bias: The selected papers report mostly positive results. Negative or inconclusive experiments are less likely to be published, leading to an optimistic view of DL-IDS performance.

Heterogeneity of setups: Differences in datasets, preprocessing, train-test splits, and metric definitions make direct numerical comparison across studies unreliable. This review therefore focuses on qualitative trends rather than exact ranking of models.

Rapid evolution: DL-IDS is a fast-moving area. New architectures (for example graph-based models, diffusion models, or large pre-trained sequence transformers) and new datasets are being proposed frequently. The conclusions drawn here may need revision as the field evolves.

VIII. Conclusion

This paper presented a focused Systematic Literature Review of deep learning methods for intrusion detection systems, based on eight representative studies published between 2018 and 2025. Two survey papers [1], [2] were used to frame the overall landscape, while five primary works provided concrete examples of CNN-based [3], LSTM-based [4], Transformer-based [5], autoencoder-based [6], and GAN-augmented [7] IDS designs, together with the CICIDS2017 dataset [8].

Regarding RQ1, the review shows that DNNs, CNNs, LSTMs, autoencoders, GANs, and Transformers are the main DL architectures applied to IDS. For RQ2, supervised CNN/LSTM/Transformer models achieve the strongest benchmark performance, while autoencoders and GANs help in representation learning and in addressing class imbalance. For RQ3, NSL-KDD, UNSW-NB15, and CICIDS2017 are the most commonly used datasets, with accuracy, precision, recall, and F1-score as standard metrics. For RQ4, persistent challenges include class imbalance, concept drift, adversarial robustness, dataset realism, interpretability, and deployment constraints.

Future work should move beyond static single-dataset benchmarks and combine:

- unsupervised and self-supervised representation learning;
- GAN or other generative augmentation for rare attacks;
- attention-based or graph-based architectures for complex dependencies;
- cross-dataset and longitudinal evaluations;
- and integrated explainability and human-in-the-loop feedback.

Only by addressing these aspects jointly can deep learning-based IDS progress from promising experimental prototypes to robust, trustworthy, and widely deployed security components.

References

[1] J. Lansky, S. Ali, M. Mohammadi, and M. K. Majeed, “Deep learning-based intrusion detection systems: A systematic review,” *IEEE Access*, vol. 9, pp. 101 574–101 599, 2021.

[2] Y. Zhang, R. C. Muniyandi, and F. Qamar, “A review of deep learning applications in intrusion detection systems: Overcoming challenges in spatiotemporal feature extraction and data imbalance,” *Applied Sciences*, vol. 15, no. 3, p. 1552, 2025.

[3] S. Vinayakumar, K. Soman, and P. Poornachandran, “Deep neural network based intrusion detection system,” in *IEEE International Conference on Signal Processing*, 2018, pp. 322–337.

[4] Y. Kim and W. Park, “Lstm-based intrusion detection system for industrial cyber-physical systems,” *Sensors*, vol. 20, no. 18, pp. 1–17, 2020.

[5] M. Peng and H. Yan, “An intrusion detection system based on transformer and feature embedding,” *IEEE Access*, vol. 9, pp. 157 454–157 465, 2021.

[6] X. Li, J. Li, and Q. Li, “Building auto-encoder intrusion detection system based on random forest,” *Computers & Security*, vol. 97, p. 101949, 2020.

[7] Y. N. Rao, P. K. S. Rao, B. Padmaja, and M. P. Suma, “An imbalanced generative adversarial network approach for network intrusion detection,” *Sensors*, vol. 23, no. 1, p. 550, 2023.

[8] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward generating a new intrusion detection dataset and intrusion traffic characterization,” in *ICISSP*, 2018, pp. 108–116.