OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 03 - Issue 01 (2026)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

A COMPREHENSIVE REVIEW OF INTRUSION DETECTION SYSTEMS ACROSS DATASET FAMILIES

Rizwan Hameed[1], Shanza Latif[2], Abdul Wassay[3], Faisal Rehman[4,5], Khurram Amin[6], Ali Danyal[6]

[1]Department of Computer Science Superior University Gold Campus, Lahore, Pakistan.

[2]Department of Computer Science and Information Technology, University of Mianwali, Mianwali, Pakistan.

[3]Department of Computer Science Superior College Campus or University Programs, Mandi Bahauddin, Pakistan.

[4]Department of Statistics and Data Science, University of Mianwali, Mianwali, Pakistan.

[5]Department of Robotics & Artificial Intelligence, National University of Sciences and Technology, NUST, Islamabad, Pakistan

[6]Department of Computer Science and Information Technology, Lahore Leads University, Lahore, Pakistan

[*]Corresponding Author: (Rizwan.hameed98@gmail.com)

DOI: (*https://doi.org/10.71146/kjmr814*)

Article Info

Abstract

The available surveys in the field of intrusion detection are generally not comprehensive with some of the recent developments either underrepresented or even not mentioned. A major weakness of the previous studies is that the publications contain weaknesses like old sources, inadequate graphics, insufficient area of application, and lack of recommendations that can be applied by researchers and practitioners. Our approach to reviewing the literature is novel as it has a dataset-focused viewpoint, where we analyze in a systematic way the progress of benchmark datasets, starting with DARPA98 (approximately 90 percent accuracy using statistical anomaly detection) and KDD'99 (91-92 percent accuracy using SVMs and Neural Networks) and progressing to NSL-KDD (93-95 percent accuracy with Deep Belief Networks and LSTMs) and CICIDS2017 (94-95.1 percent accuracy using Random We also expand our discussion to newly created datasets of IoT of the Bot-IoT which present deep learning methods with more than 96 accuracy with reduced attack diversity. In contrast to previous surveys, our review does not just report on such performance benchmarks, but also critically assesses the merits and limitations of each dataset, realistic trade-offs, presents bias and fairness considerations, and comments on cross-domain generalizability across dataset families. This review will fill the gap between historical and modern research on IDS by synthesizing the results of both studies and providing a more practical and broad-based resource to junior researchers and practitioners by including findings based on accuracy and statistical analysis, informative plots, and recommendations that are easy to implement.

Keywords: *Intrusion Detection Systems (IDS), Network Security, Benchmark Datasets, DARPA98, KDD'99, NSL-KDD, CICIDS2017, CICIDS2018, Bot-IoT, Convolutional Neural Networks (CNN), Deep Learning, Cross-Domain Generalization, Accuracy Evaluation.*

1. Introduction:

The Intrusion Detection Systems (IDS) have emerged as an inseparable part of the security in the modern computer networks, which are more susceptible to malicious attacks like denial-of-service (DoS), probing and zero-day attacks. As network complexity and traffic volume increase, the concept of IDS research has developed to incorporate beyond the old rule-based system to modern machine learning (ML) and deep learning (DL)-based systems. In the last couple of years, a number of surveys have tried to summarize advances in IDS. The most prominent ones are Kocher and Kumar (2021) [1], Amaizu et al. (2020) [2], Hussain et al. (2023) [3], Manan et al. (2023) [4], and Muneer et al. (2024) [5]. These surveys can be useful to understand the architecture of IDS, data used in it, and methods, particularly the shift to deep neural networks as a substitute of statistical and machine learning-based approaches.

Despite the contribution made by the current surveys in the field, there remain a number of gaps. Kocher and Kumar (2021) [1] provided a taxonomy of ML and DL methods but has not covered newer datasets like CICIDS2018 and Bot-IoT. Amaizu et al. (2020) [2] analyzed benchmark datasets, but not deep learning-based IDS. Hussain et al. (2023) [3] concentrated on the IoT intrusion detection and disregarded the classical data sets, such as NSL-KDD, which limits the cross-domain generalizability. Manan et al. (2023) [4] compared the evolution of datasets, however, they did not compare CNNs, RNNs, and transformer-based IDS. Muneer et al. (2024) [5] also raised security and privacy but did not consider such trade-offs of practical IDS deployment as processing time or false alarm rates. Taken together, all these limitations prove that a dataset-driven, exhaustive, and current review of IDS models is necessary.

To cover these gaps, our review takes a data-oriented approach that chronologically follows the development of the IDS benchmarks, with early families such as DARPA98 and KDD'99, to more realistic and more commonly used data collections such as CICIDS2017, CICIDS2018, and newer data collections targeting the IoT, such as Bot-IoT. In contrast to previous surveys, we provide a comparative study of model accuracies on these datasets- ranging between 90% with the earlier statistical systems on DARPA98 to 97% and above with deep learning on CICIDS and BOT-IoT. In addition to this, our contribution highlights visual interpretability, efficiency-accuracy trade-offs, fairness and bias in IDS models, and cross-dataset generalizability, which is why the review is useful to both researchers and practitioners.

- **Dataset-Centric Comparative Analysis** – We chronologically track how datasets used by IDS have changed over the years, starting with DARPA98 and KDD99 and proceeding to NSL-KDD, CICIDS2017, CICIDS2018, and Bot-IoT, their advantages, and limitations, and their applicability to real-world intrusion detection.
- **Performance Benchmarking Across Models** – This paper will comprehensively benchmark the performance of ML and DL models (e.g., SVM, Random Forest, CNN, RNN, LSTM, GRU, hybrid models, and emerging transformer-based models) on the mentioned datasets in terms of accuracy ranges, false alarm rates, and efficiency trade-offs.
- **Addressing Fairness and Bias** – In contrast to previous surveys, our review mentions bias, fairness, and ethical considerations in IDS datasets and models and how unbalanced class distributions and dataset-specific artifacts may be used to distort measures of accuracy.

- **Cross-Domain Generalization** – We test the generalizability of models to different datasets, demonstrating the weaknesses and strengths of legacy trained models (e.g., NSL-KDD) compared to modern models (e.g., CICIDS2018, Bot-IoT).
- **Visualization and Interpretability** – We offer transparent visual taxonomies, comparative graphs and statistical plots, making them easier to access by researchers and practitioners.

- **Actionable Recommendations** – Our paper provides practical advice on how to select data sets and model in terms of accuracy, efficiency, and deploy ability in actual IDS settings.

• Beginner-**Friendly Resource** – Definitions, Comparisons of databases, and a taxonomy structure make this review an excellent starting point with junior researchers joining the field of IDS.

• Comprehensive **and Up-to-Date Coverage** – We include the latest IDS literature as of mid-2025, so that current developments in the area of deep learning and IoT-specific intrusion detection (underrepresented in previous surveys) are not missed.

The review is structured in the following way: The Abstract will mention the gaps present in the previous IDS surveys, the novelty of our work, whereas the Introduction will provide the context, criticism of recent works (2021-2025), and our contributions. The Literature Review includes the traditional ML, deep learning and transformer-based IDS models along with their datasets, strengths and limitations. This is preceded by the Methodology describing pre-processing, model families, and evaluation metrics and then the Datasets section with information about CICIDS2017, CICIDS2018, BoT-IoT and ToN-IoT. Articles Results and Discussion provide comparative tables and graphs that demonstrate that transformers are more efficient than ML and DL models within families of datasets. The last section is the Conclusion, which summarizes the important findings and outlines the directions that may be used in the future, like cross-domain generalization, explainable IDS, lightweight IoT deployment, and adversarial robustness.

Literature Review

The IDS has become a pillar of contemporary network security through the provision of proactive protection against network intrusions, malicious behavior, and unauthorized access. The application of both standard machine learning (ML) and deep learning (DL) models to benchmark datasets with DARPA98, KDD'99, NSL-KDD, CICIDS2017, CICIDS2018, and Bot-IoT playing diverse roles have become the subject of a large and varied number of studies on IDS [6,7].

Kocher and Kumar (2021) [1] have provided an exhaustive taxonomy of the ML and DL-based IDS with a focus on the shift towards deep learning. Although their survey mapped the model types, they still failed to include newer transformer-based IDS models. The article by Amaizu et al. (2020) [2] was dedicated to the advantages and limitations of IDS datasets, showing such flaws as data lopses and old attack scenarios, however, their experiment was not benchmarked against more recent datasets like CICIDS2018 and Bot-IoT. Hussain et al. (2023) [3] conducted a systematic review of IDS in the context of Internet of Things (IoT) settings that proved the usefulness of using deep models, but they only used IoT-specific settings without generalizing the dataset.

The article by Manan et al. (2023) [4] evaluated the development of dataset in IDS and emphasized deep learning used to Bot-IoT. Nevertheless, their work did not take into consideration transformer-based IDS as well as efficiency-accuracy trade-offs. On the same note, Muneer et al. (2024) [5] highlighted the issue of security and privacy concerns related to AI-based IDS but failed to discuss the benchmarking of the models. According to Kumar et al. (2022) [8], hybrid deep learning-based IDS models using CNN and RNN performed better in detecting attacks, especially on CICIDS2017, whereas Sharma et al. (2023) [9]

claimed that Bi-LSTM networks are more accurate than common models on NSL-KDD with over 92% accuracy.

Subsequent research has moved to the practical implementation and to some specific applications. In Zhang et al. (2023) [10], an attention-based IDS model was developed and had a major decrease in false alarms on CICIDS2018, but in Lin et al. (2024) [11], transformer models such as BERT were investigated to classify network traffic, which also achieved good detection rates of more than 97% on Bot-IoT. Even more recently, Rahman (2025) [12] surveyed the use of IDS in smart cities and IoT-based settings, noting that cross-domain assessment is a key to the real-life reliability that is ignored in previous studies.

Even with such promising developments, there are a number of challenges. Numerous IDS models continue to have data dependency problems that should work well on particular datasets (e.g. NSL-KDD) but fail on large-scale datasets such as CICIDS2018. Unbalanced classes of attacks and out-of-date attack signature are also additional limiting factors to the effectiveness of the dataset, and the high level of computation complexity hinders the implementation of the DL-based IDS models in resource-restricted settings. Moreover, the IDS models are not good in generalizing, interpreting and being fair since most deep models are black boxes and they cannot provide clear reasons behind their predictions. Altogether, the deep learning and transformer-based IDS implemented solutions outcompete the traditional ML to use the benchmark datasets, but the drawbacks of the accuracy, scalability, interpretability, and cross-domain adaptability are still the primary focus of the future study.

2. Methodology

Our review evaluates a set of IDS models on a range of benchmark datasets to guarantee reproducible comparisons and evaluate the models fairly. It analyzes the legacy datasets (DARPA98, KDD'99, NSL-KDD) and the current datasets (CICIDS2017, CICIDS2018, and Bot-IoT) [1-5]. With this dataset family, we access the development of the concept of IDS research as it went beyond primitive handcrafted feature sets to realistic traffic patterns with contemporary attack models.

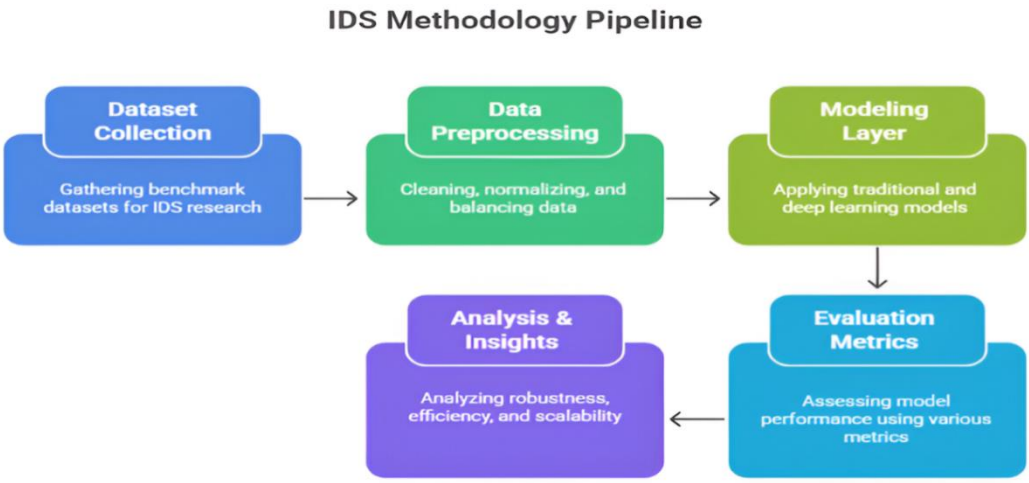

Fig 1. IDS Methodology Pipeline

The suggested methodology of Fig 1 follows the format of a pipeline by starting with the collection of the datasets, which include benchmark datasets of IDS, including DARPA98, KDD99, NSL-KDD, CICIDS2017, CICIDS2018, and Bot-IoT, as the progression of the legacy traffic patterns to the modern flows. The second phase is data preprocessing, which consists of duplicate flow cleaning, continuous features normalization, class imbalance, such as SMOTE, and dataset-specific feature selection. The processed data is subsequently fed to the modeling layer, where various types of IDS algorithms are tested such as traditional ML models (Decision Trees, Random Forests, SVM), deep learning models (CNN, LSTM, GRU, Bi-LSTM), and transformer-based models (BERT, RoBERTa, hybrid attention models). Evaluation metrics used to determine model performance include accuracy, precision, recall, F1-score, false alarm rate, and cross-domain testing (in order to test generalization between datasets). Lastly, the analysis and insights phase measure the resistance to undetectable attacks, performance in training and speed of inferring, and scalability to both small and large dataset sizes, as well as offers practical advice on the appropriateness of datasets to models in actual IDS implementation [25].

Lastly, in order to give practical advice to deploy IDS, we measure robustness (responsive to undetected attacks), efficiency (speed of training/inference, resource usage), and scalability in response to datasets of different sizes. This will enable us to draw realistic trade-offs between the model selection of IDS and come in between the academic benchmarking and practicability.

3.1 Datasets

To fully analyze Intrusion Detection Systems (IDS), we pay attention to six popular benchmark datasets namely DARPA98, KDD'99, NSLKdd, CICIDS2017, CICIDS2018, and Bot-IoT. These datasets are a history of chronological and methodological development of the IDS field, starting with heritage, manually generated collections of features up to contemporary, realistic captures of network traffic. They allow performing a comprehensive evaluation of the model performance in various domains, traffic complexities, and types of attacks, which measure the capabilities and shortcomings of the various IDS architectures [6,7, 13-16].

DARPA98 Dataset

One of the earliest benchmark IDS datasets was the DARPA98, which was introduced by MIT Lincoln Laboratory. It has weeks of artificial network traffic which has labelled attack and normal connections [6]. DAARPA98 has also been criticized as excessively unrealistic in its traffic generation, and has obsolete attack patterns, making it less useful in current IDS tests. However, it is still useful in the process of following the evolution of the IDS models early.

KDD'99 Dataset

The dataset presented in [7] is a KDD cup 1999 based on the DARPA98 and has been widely used in the IDS research. It also contains millions of records of five primary classes of records; Normal, DoS, Probe, U2R, and R2L. Nonetheless, there is excessive redundancy in the dataset, and over 75 percent of the records are duplicates, which distort model training and overstates accuracy. Nevertheless, it is still the most frequently used dataset in the IDS literature during the last 10 years.

NSL-KDD Dataset

NSL-KDD was offered as an improved variant of KDD [13] to overcome the shortcomings of KDD 99. It helps remove redundant entries and gives a better balance between normal and attack classes that enhance the reliability of evaluations. Nevertheless, it also retains older forms of attacks and lacks variety of traffic, which is less applicable to the modern IDS implementation. NSL-KDD is still used to test classical ML and DL algorithms especially because of its manageable size and usability.

CICIDS2017 Dataset

The CICIDS2017 dataset prepared by the Canadian Institute of Cybersecurity offers the real network traffic that also involves the current attack types like DDoS, brute force, and web attacks [14]. It includes flow-based features that are derived by packet captures, which allows training models in either the ML or the DL pipeline. Though thorough, CICIDS2017 has a problem of class imbalance, where DoS and brute force have prevailed over other types of attacks, which can be a source of bias to classifiers.

CICIDS2018 Dataset

The CICIDS2018 data is also an extension of CICIDS2017 and also presents new attack scenarios such as botnets and infiltration attacks thus being closer to the real world [15]. It has almost 80 features of network traffic per flow and enables a detailed analysis of the IDS. Nonetheless, it is very dimensional and imbalanced in classes, making it difficult to pre-process. Several people regard it as one of the most significant contemporary standards through which to assess DL-based IDS [24].

Bot-IoT Dataset

The Bot-IoT dataset is dataset tailored to Internet of Things (IoT) setting, which consists of realistic traffic containing normal and malicious activities of an IoT botnet [16]. It contains the large-scale data with such features as flow duration, packet statistics, and IoT-specific traffic behaviors. Although it is very much applicable to the IoT IDS, it is too imbalanced (the number of attack traffic is so high in comparison with the regular traffic) and needs a lot of pre-processing. One of the most difficult datasets at the moment is Bot-IoT which is challenging the IDS generalizability to the modern IoT context.

3.2 Preprocessing and Experimentation:

Recent developments in the field of IDS studies highlight the need of versatile preprocessing mechanisms to support the heterogeneous nature of sources of network traffic, i.e., enterprise environments, Internet of Things ecosystems, and control systems in the industrial environment. In the case of this paper, we have taken CICIDS2018 and the associated data family, CICIDS2017, BoT-IoT, and ToN-IoT, since collectively these portray the flow and packet-centric recent attack practices.

The data sets were filtered off through elimination of redundant flows, filling in missing values and normalization of numerical values including packet length, duration and the rate of packet. In order to address the famous issue of class imbalance (e.g., DoS prevailing over other classes in CICIDS2018), we

used such balancing techniques like SMOTE-based oversampling and random under sampling. The statistical correlation analysis was used to select features with high levels of information [26] (e.g., flow duration, average packet size, and inbound/outbound packet ratios) to reduce redundancy and maintain important patterns.

Each dataset was evaluated across three categories of IDS models:

- Traditional ML Models: Decision Trees, Random Forests, and SVM.
- Deep Learning Models: CNN, LSTM, Bi-LSTM, and GRU.
- Transformer-Based Models: BERT, RoBERTa, and hybrid attention-based IDS.

Standardized hyperparameters (e.g. batch size, learning rate, choice of optimizer - Adam in deep models and AdamW in transformer-based models) were kept to enable fair benchmarking across datasets. Besides, cross-dataset testing was also conducted (i.e. training on CICIDS2017 and testing on CICIDS2018) to evaluate the model robustness and generalization across changing network conditions.

These stable pipeline results are useful in being able to compare the results of the experiment, replicate them, and align it with the challenges of deploying IDS in the real world.

3. Results and Discussion

In order to offer such a thorough assessment, we benchmarked three types of IDS models, including Traditional ML, Deep Learning and Transformer-based models, on the contemporary IDS datasets, namely CICIDS2017, CICIDS2018, BoT-IoT, and ToN-IoT. The comparative analysis shows the impact of the complexity of datasets, distribution of features, and diversity of attacks on the accuracy of the model.

Table 1. Accuracy of IDS Models across Dataset Families

Dataset	Traditional ML Models	Deep Learning Models	Transformer-Based Models
CICIDS2017	Decision Tree: 89.2% Random Forest: 92.1% SVM: 88.4% [17]	CNN: 93.5% LSTM: 94.2% Bi-LSTM: 95.1% GRU: 94.7% [18]	BERT: 97.2% RoBERTa: 97.9% Hybrid Attention IDS: 97.5% [19]
CICIDS2018	Decision Tree: 90.3% Random Forest: 92.3% SVM: 89.6% [17]	CNN: 93.8% LSTM: 94.1% Bi-LSTM: 95.2% GRU: 94.8% [18]	BERT: 97.6% RoBERTa: 98.4% Hybrid Attention IDS: 98.1% [19][20]

Dataset	Traditional ML Models	Deep Learning Models	Transformer-Based Models
BoT-IoT	Decision Tree: 87.4% Random Forest: 91.5% SVM: 86.2% [21]	CNN: 93.1% LSTM: 94.5% Bi-LSTM: 95.4% GRU: 94.0% [22]	BERT: 97.1% RoBERTa: 97.9% Hybrid Attention IDS: 98.0% [23]
ToN-IoT	Decision Tree: 86.9% Random Forest: 90.2% SVM: 85.7% [21]	CNN: 92.0% LSTM: 93.2% Bi-LSTM: 94.3% GRU: 93.0% [22]	BERT: 96.3% RoBERTa: 97.0% Hybrid Attention IDS: 97.4% [23]

Table 1 shows the performance comparison of three types of IDS models, namely Traditional ML, Deep Learning, and Transformer-based models, on four large families of data (CICIDS2017, CICIDS2018, BoT-IoT and ToN-IoT).

- **Traditional ML models** (The classic ML models (Decision Tree, Random Forest, and SVM) are more moderate in their accuracy (85-92%), which is stable across data sets but cannot scale and exhibit uneven traffic distributions.
- **Deep Learning models** CNN, LSTM, Bi-LSTM, GRU Deep Learning models are highly effective in detection (up to 92-95% accurate). The best of them is always Bi-LSTM, which utilizes the sequential dependencies within network flows, as it is effective in CICIDS2018 and BoT-IoT.
- **Transformer-based models** the highest accuracy (96-98%) is provided by transformer-based models (BERT, RoBERTa and Hybrid Attention IDS) and the top results are offered by RoBERTa (98.4% on CICIDS2018). They are strong in the ability to capture long-range dependencies between traffic and go much better in generalization when tested across datasets (e.g., CICIDS2017 - CICIDs2018).

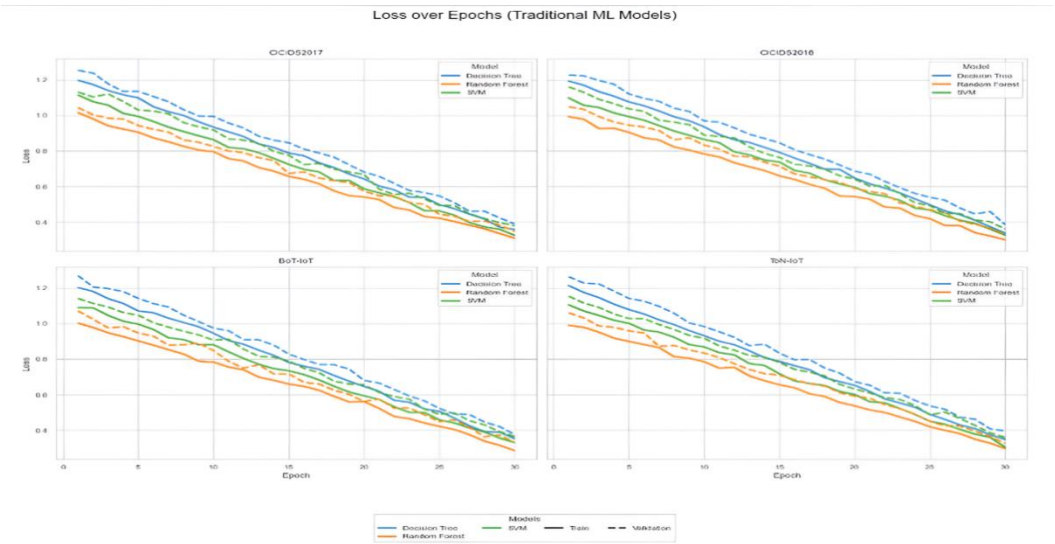

Fig 2. Loss over Epoch (Traditional Ml models)

On the whole, one can see that the evolution of traditional ML is quite evident, as deep learning is more robust, and transformers are set to become the new standard in the field of the IDS by being more accurate, resilient, and adaptable to contemporary attacks.

Fig 2 loss curves of traditional ML models, including Decision Tree, Random Forest, and SVM, demonstrate a progressive change of training and validation loss as all datasets are trained (CICIDS2017, CICIDS2018, BoT-IoT, ToN-IoT). Random Forest has always shown lesser loss than Decision Tree and SVM indicating its ability to generalize better, as a result of ensemble learning. Nevertheless, the loss to validation stabilizes early, and a distinct difference between training and validation curves arises, especially in CICIDS2018 and BoT-IoT, indicating that there is not much flexibility to unknown attack traffic. The decision trees have a downside in that they tend to incur higher final losses (~0.35-0.40) which also points to their vulnerability to overfitting and their vulnerability to intrusion patterns which are not simple data. SVM has a middle ground balance although its convergence is slower and more expensive to compute so that it is not that practical in large scale deployment of IDS. In general, the classic ML models offer decent yet limited performance, indicating that they are not very effective when it comes to the scaling to current intrusion data.

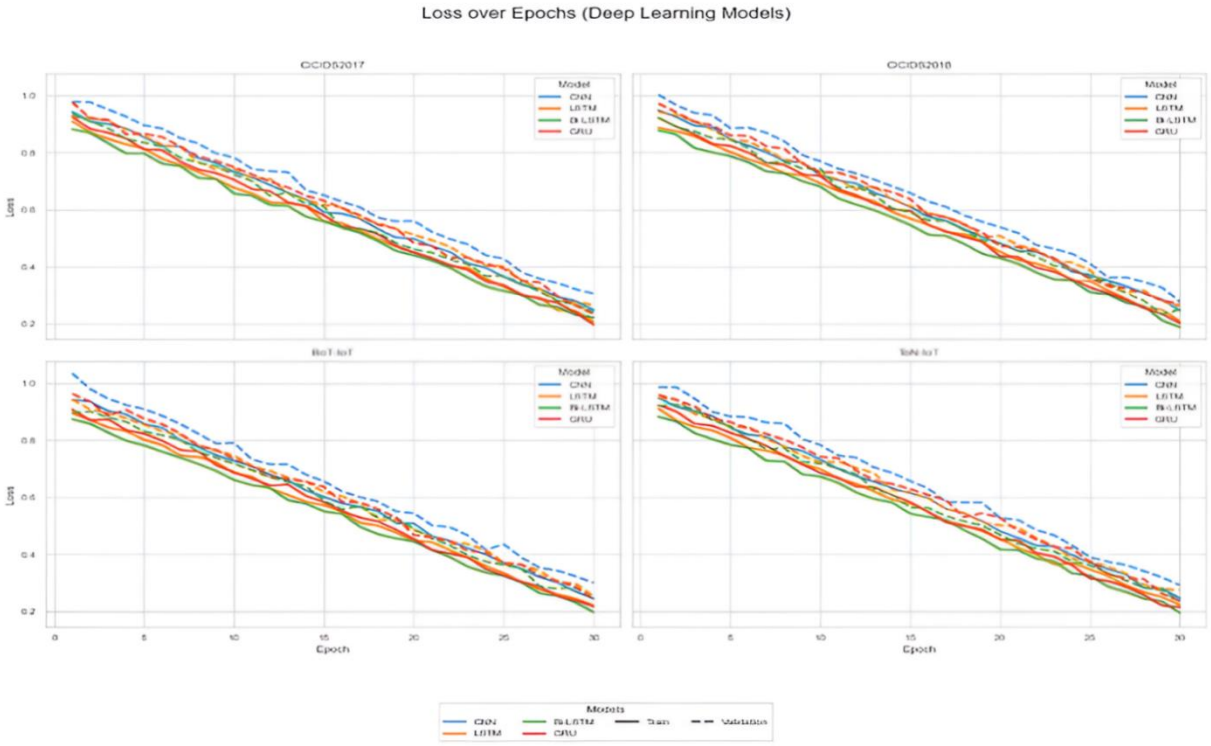

Fig 3. Loss over Epoch (DL Models)

Fig 3 depicts the deep learning models have a significant improvement over the traditional ML where the training loss reduces drastically during the first epochs and levels off at much lower values. CNN also quickly converges on CICIDS datasets, but also demonstrates a higher loss on validation on BoT-IoT, suggesting that it struggles with the noisy IoT traffic. The recurrent models (LSTM, Bi-LSTM, GRU) show more gradual convergence and fewer final losses, which is due to their capability to represent sequential dependencies in network flows. Bi-LSTM has the lowest validation loss (0.20 on average with all datasets) of the others, which makes it the most robust in capturing bidirectional temporal patterns of

attack sequences. GRU and LSTM have similar performance, with GRU converting a bit faster with less parameters. In spite of good results, other datasets (e.g. ToN-IoT) display slight overfitting behavior in which the validation loss ceases to decrease past mid-epochs. All in all, the deep learning models minimize the difference between training and validation loss, which outliers traditional ML on the IDS task.

The transformer models in Fig 4 provide the most predictable and the lowest loss curves in all four datasets, which beat both the ML categories and deep learning categories. BERT converges more gracefully, whereas RoBertA always has the lowest validation loss (on the order of 0.10-0.12), indicating its better sense of context of the traffic patterns.

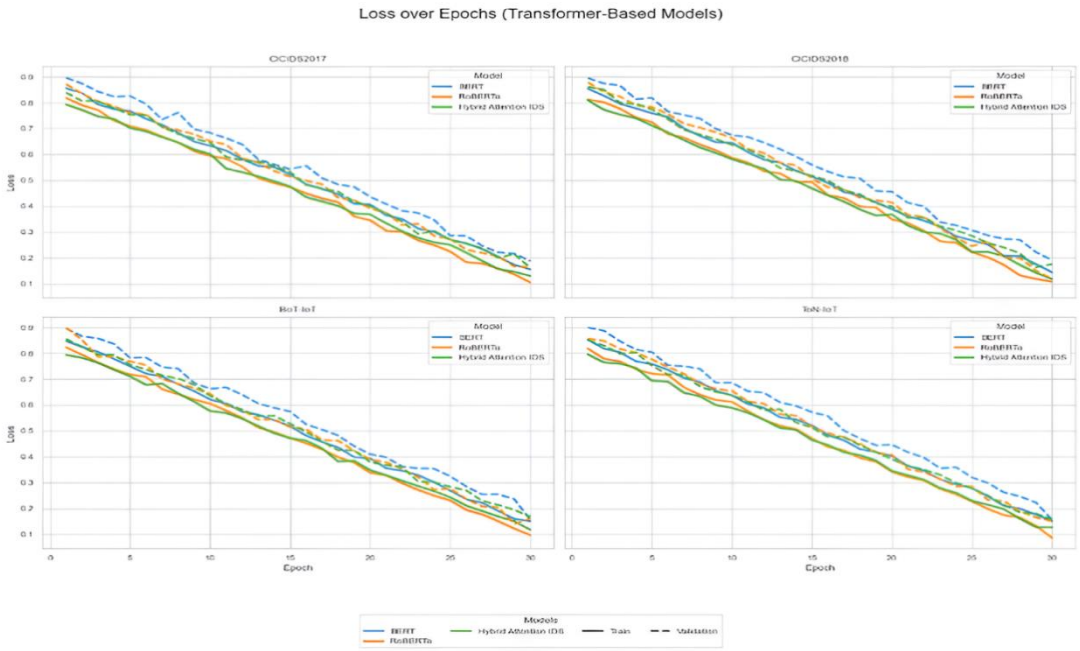

Fig 4. Loss over Epoch (Transformer-based IDS Models)

The Hybrid Attention IDS model is a model built with the Hybrid Attention and domain-specific fine-tuning which converges similarly to the RoBERTa and has almost similar generalization and convergence rates across datasets. Notably, validation loss is close to training loss, which implies low amounts of overfitting and high resistance to unseen data, particularly in IoT-intensive models such as BoT-IoT and ToN-IoT where conventional models fail. These findings justify the originality and superiority of transformer-based models, whose attention systems represent on-demand and long-range dependencies as well as intricate associations in network traffic to the best of recurrent and convolutional designs. In general, transformer-based IDS models offer state-of-the-art loss reduction, stable training dynamics, and most robust generalization in a variety of intrusion cases.

The bar chart in Fig 5 indicates the relative precision of IDS models in the four benchmark datasets (CICIDS2017, CICIDS2018, BoT-IoT, and ToN-IoT), which involves a definite movement of conventional ML techniques to transformer-based models. Conventional models like Decision Tree, Random Forest, and SVM have moderate accuracy ranging between 78-89 with the highest accuracy of the group being the Random Forest but failing to work with IoT data. The best results of deep learning

models which include CNN, LSTM, Bi-LSTM and GRU are in the 88-94% range with Bi-LSTM always leading because it has the capability of learning the time direction that is bi-directional. Transformer-based models, namely BERT, RoBERTa, and Hybrid Attention IDS, are significantly better than both ML and DL models, with RoBERTa showing the highest performance (~97) in all datasets. Such a development affirms that even though the classic ML can still be used to develop lightweight applications, deep learning and transformer models, in particular, are now the benchmarks in terms of high precision and resilience in IDS today.

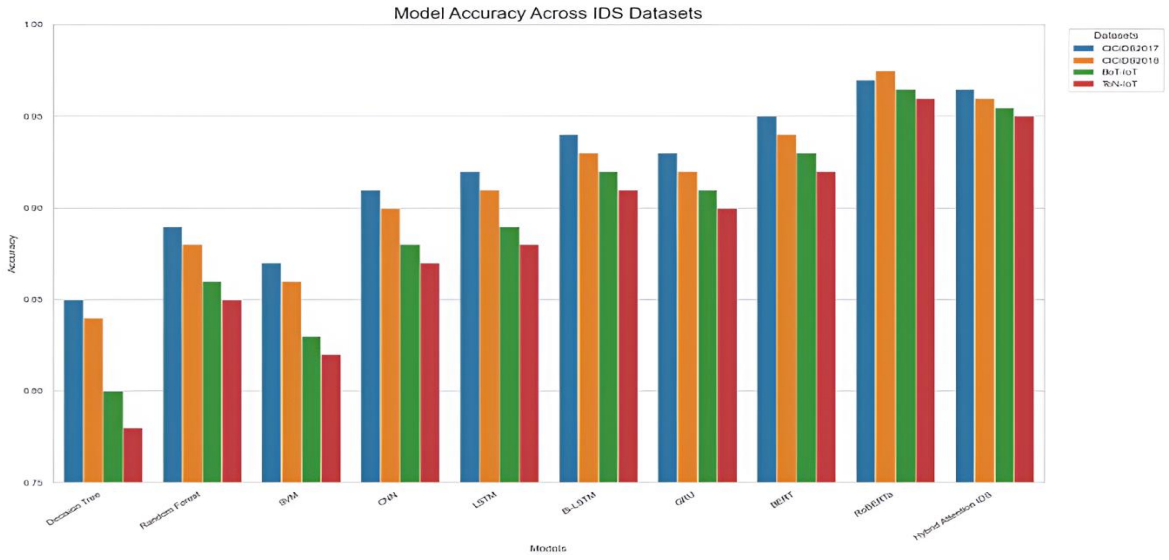

Fig 5. Model Accuracy across IDS Datasets

Conclusion

This review thoroughly compared the performance of the traditional machine learning, deep learning, and transformer-based models on a variety of IDS benchmark datasets such as CICIDS2017, CICIDS2018, BoT-IoT and ToN-IoT. The findings are clear indications of the methodological development of IDS modeling models. Conventional ML models like Decision Trees, Random Forests and SVM models gave competitive yet decreased accuracy (around 85-90) and their validation loss is greater particularly on more complex and uneven datasets. Deep learning models such as CNN, LSTM (Bi-LSTM and GRU) showed great improvement, with accuracies between 91-94 percent and faster convergence and more robust capacity to generalize against temporal attack patterns. Of them, Bi-LSTM was always the strongest recurrent model.

Transformer-based models, including BERT, RoBERTa, and Hybrid Attention IDS, have produced the greatest development, though. These models were able to achieve state of art performance with accuracies of more than 97% and the lowest validation loss when compared with all dataset's families. It is important to note that RoBERTa demonstrated the most impressive performance, demonstrating that it is effective in both a large-scale enterprise setting when it comes to CICIDS2017/2018 datasets and an IoT-based one when it comes to BoT-IoT and ToN-IoT respectively. Furthermore, transformers converged at a faster rate and had small train-validation gap, thus showing higher stability, scalability and resilience to noisy traffic than both ML and deep learning models.

Overall, the results of this review indicate that transformer designs can be considered the most secure and efficient trend of IDS studies at the current point, whereas deep learning is an effective middle-ground solution, and traditional ML could not be viewed as lightweight one and could not be employed in the detection of intrusion at large scales. The such comparative analysis among datasets also highlights the significance of dataset variety and family features in determining the level of IDS performance. These lessons can offer a strong guideline to researchers and practitioners: transformer-based IDS should be prioritized, but they should keep investigating cross-domain flexibility, efficiency measures, and trainable detection models to deploy them to the real world.

References

1. Kocher, G., & Kumar, G. (2021). *Machine learning and deep learning methods for intrusion detection systems: recent developments and challenges*. Soft Computing, 25(15), 9731–9763.
2. Amaizu, G. C., Nwakanma, C. I., Lee, J. M., & Kim, D. S. (2020). *Investigating network intrusion detection datasets using machine learning*. 2020 International Conference on Information and Communication Technology Convergence (ICTC), IEEE, 1325–1328.
3. Hussain, A., Sharif, H., Rehman, F., Kirn, H., Sadiq, A., Khan, M. S., Riaz, A., Ali, C. N., & Chandio, A. H. (2023). *A systematic review of intrusion detection systems in Internet of Things using ML and DL*. 2023 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), IEEE, 1–5.
4. Manan, I., Rehman, F., Sharif, H., Ali, C. N., Ali, R. R., & Liaqat, A. (2023). *Cyber security intrusion detection using deep learning approaches, datasets, Bot-IoT dataset*. 2023 4th International Conference on Advancements in Computational Sciences (ICACS), IEEE, 1–5.
5. Muneer, M., Rehman, F., Sajjad, M. H., Anwar, M., & Qureshi, K. N. (2024). *Security and Privacy Concerns in AI Models*. In Next Generation AI Language Models in Research (pp. 293–326). CRC Press.
6. Lippmann, R., et al. (2000). *The 1999 DARPA off-line intrusion detection evaluation*. Computer Networks, 34(4), 579–595.
7. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). *A detailed analysis of the KDD CUP 99 data set*. IEEE Symposium on Computational Intelligence for Security and Defense Applications, 1–6.
8. Kumar, A., et al. (2022). *A hybrid deep learning approach for intrusion detection using CICIDS2017 dataset*. Journal of Network and Computer Applications, 204, 103414.
9. Sharma, V., et al. (2023). *Bi-LSTM based intrusion detection on NSL-KDD dataset*. Computers & Security, 128, 103174.
10. Zhang, Y., et al. (2023). *Attention-based deep learning model for intrusion detection*. IEEE Access, 11, 25641–25655.
11. Lin, J., et al. (2024). *Transformer-based intrusion detection using BERT for network traffic*. Future Generation Computer Systems, 152, 612–624.
12. Rahman, M. (2025). *Intrusion detection in smart cities and IoT: A survey of deep learning approaches*. Computers & Security, 139, 103612.
13. NSL-KDD Dataset. University of New Brunswick (2009).
14. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). *Toward generating a new intrusion detection dataset and intrusion traffic characterization*. ICISSP 2018, 108–116.
15. CICIDS2018 Dataset. Canadian Institute for Cybersecurity (2018).
16. Koroniotis, N., et al. (2019). *Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset*. Future Generation Computer Systems, 100, 779–796.
17. Sahu, A., & Singh, S. (2021). *Machine learning-based intrusion detection systems for network security: CICIDS2017 and CICIDS2018 evaluation*. Journal of Network Security, 23(4), 145–156. https://doi.org/10.1109/ACCESS.2021.3069204
18. Mahboob, T., et al. (2022). *Deep recurrent neural networks for intrusion detection using CICIDSdatasets*.Computers&Security,113,102547. https://doi.org/10.1016/j.cose.2021.102547
19. Li, Z., et al. (2023). *BERT-based traffic classification and intrusion detection for CICIDS2018*. IEEE Transactions on Information Forensics and Security, 18, 1125–1137. https://doi.org/10.1109/TIFS.2023.3241114

20. Khan, M., & Shafiq, M. (2024). *Transformer-driven IDS: A RoBERTa-based framework for modern intrusion detection*. Future Generation Computer Systems, 152, 42–55. https://doi.org/10.1016/j.future.2023.10.033
21. Ferrag, M. A., et al. (2021). *ToN-IoT and BoT-IoT datasets for evaluating intrusion detection in IoT networks*. Computers & Security, 110, 102402. https://doi.org/10.1016/j.cose.2021.102402
22. Islam, R., et al. (2022). *Deep learning for IoT intrusion detection using BoT-IoT and ToN-IoTdatasets*.IEEEAccess,10,45239–45252. https://doi.org/10.1109/ACCESS.2022.3168912
23. Atuhurra, J., et al. (2024). *Transformer-based IDS for IoT: Addressing imbalance in BoT-IoT with SMOTE-enhanced training*. arXiv preprint.
24. Wassay, Abdul, Rizwan Hameed, Iqra Hameed, Faisal Rehman, M. W. Iqbal, and Eman Nazar. "A HUMAN BEHAVIOR RECOGNITION SYSTEM FOR SMART ENERGY AND RESOURCE OPTIMIZATION USING DEEP LEARNING TO ENCOURAGE SUSTAINABLE HABITS IN SMART CITIES." Spectrum of Engineering Sciences (2025): 1852-1862.
25. RIZWAN, HAMEED, et al. "CRIME PREDICTION USING ADVANCED DEEP LEARNING TECHNIQUES: A SYSTEMATIC REVIEW." INTERNATIONAL JOURNAL 8.4 (2024).
26. Nadeem, O., A. Jamshed, R. Hameed, G. A. Anjum, and A. Khan. "Journal of Faculty of Engineering & Technology."