AI and Machine Learning Advances
Volume 2, Issue 1
ISSN: 3067-3216

Evaluating the Evolution and Efficacy of AI-Driven Intrusion Detection Systems Against Zero-Day Attacks

[1]*Krishnam Nimmala
[1]Information Technology/software engineering, Independent Researcher.

Abstract

There is a high rate of cyber threats development and the use of the zero-day attack, which presents a great challenge to the conventional intrusion detection system (IDS). This paper compares the performance and the generalization ability of the AI-based IDS models on two benchmark datasets (NSL-KDD (legacy) and CIC-IDS2017 (modern)) datasets. We present comparisons of Random Forest, XGBoost, and Support Vector Machine in supervised and zero-day simulation scenarios in a leave-one-attack-out set up. Supervised performance has been observed to be almost perfect on both datasets, with recall and ROC-AUC scores being close to 0.999 with the tree-based models. Nevertheless, zero-day analysis demonstrates significant performance reduction, and a drop to about 68 and 58 percent on NSL-KDD and CIC-IDS2017, respectively. These results demonstrate that there is a severe disparity between controlled precision and actual generalization in the real world. The findings show that AI-based IDS models are effective in detecting known attacks but have poor zero-day resiliency, and it is important to note that more generalized and adaptable intrusion detection systems should be designed.

Keywords: Artificial Intelligence (AI), Intrusion Detection Systems (IDS), Zero-Day Attacks, Machine Learning Models, Cybersecurity, Random Forest, XGBoost.

1. Introduction

1.1 Background and Motivation

The fast-evolving nature of cyber threats has drastically altered the security environment of the contemporary digital infrastructures. As cloud computing spreads, as do Internet of Things (IoT) ecosystems and large-scale distributed networks, organizations are confronted by increasingly intelligent and resilient attacks. In this list of threats, zero-day vulnerabilities are one of the most significant security threats, since they take advantage of the unknown vulnerabilities that have not been patched or even countermeasures developed. The traditional signature-based intrusion detection systems (IDS) is a priori limited to the detection of such attacks since it is limited to a set of established patterns based on threats that have been known. Consequently, they find it difficult to detect new attack patterns that are not shown in the signatures.

A recent study has highlighted the increased significance of artificial intelligence (AI) and machine learning (ML) methods to reducing such constraints. Zero-day detectors that use machine learning have proved to be capable of studying behavioral patterns and identifying anomalies outside of the preset rule sets (Guo, 2022). Hybrid frameworks of intrusion detection based on the combination of ensemble learning and deep architecture have demonstrated encouraging outcomes in enhancing the detection resilience (Dai et al., 2024). In addition, recent developments such as deep learning models, such as long short-term memory (LSTM) networks and transfer learning models, have further increased the ability of IDS models to model intricate traffic dynamics (Kansal, 2025; Rodríguez et al., 2022). These innovations highlight the shift of reactive and signature-based detection towards adaptive mechanisms of intrusion detection using AI.

1.2 Problem Statement

Although there has been a remarkable advancement in AI-based IDS studies, there has been a consistent challenge in identifying invisible or zero-day attacks. In a well-supervised environment, many machine learning models may perform almost perfectly, but these results are likely to be optimized to known attack distributions, and not necessarily to the actual generalization ability. The tendency to overfit benchmark data sets can lead to artificial accuracy measures that are not applicable in practice.

Krishnam Nimmala
Information Technology/software engineering, Independent Researcher.
Email: meetnimmala@gmail.com

Received: 11-Feb-2026 Revised: 25-Feb-2026 Accepted: 8-March-2026

AI and Machine Learning Advances
Volume 2, Issue 1
ISSN: 3067-3216

The current research will often test models on one dataset, which restricts the information on cross-dataset generalization. Contemporary attack settings are quite different than the simulated datasets of the past, and live deployed models that are trained using a single traffic distribution might not work when the model is applied to a heterogeneous network setting. The new literature has recognized the growing sophistication of AI-related cyber threats and the necessity of flexible detection systems (Alansary et al., 2025; Hashim et al., 2025). There are few systematic reviews that measure the performance of generalization on datasets. As a result, it is urgently necessary to determine the true value of AI-powered models of IDS in enhancing zero-day detection or improve performance when operating in controlled experimental settings.

1.3 Research Gap

An overview of the modern literature indicates that there are a number of significant vacua. To begin with, the majority of the studies employ one benchmark dataset in assessing zero-day detection efficacy (Dai et al., 2024; Guo, 2022). Although such assessments offer controlled experimentality, they are inadequate in regard to cross-environment robustness. Second, very little literature contains explicit comparisons of legacy datasets, e.g. NSL-KDD, and contemporary flow-based datasets modeling realistic enterprise traffic patterns. This difference is important, because the evolution of the dataset affects the aspect of feature representation and model behavior to a considerable degree.

Third, methodologies of the zero-day simulation are still not consistent. Anomaly-based detection strategies or hybrid or transfer-learning strategies are used in some works, and the standardized leave-one-attack-out strategies are not constituted in others (Karn et al., 2025; Rodríguez et al., 2022). Systematic benchmarking under the same evaluation standards has not been performed extensively, but federated and rare-attack detection strategies have been suggested to improve robustness (Abhijit et al., 2025; Lin, 2025). To develop research in IDS, there is a need to have a structured, reproducible zero-day simulation framework that will allow cross-model and cross-dataset comparisons.

1.4 Research Objectives

This study aims to address the identified gaps through a structured comparative analysis of traditional machine learning and AI-driven intrusion detection models. The objectives are:

- To determine the performance of ensemble-based and deep learning IDS models.
- To compare the performance of detection on two datasets NSL-KDD (legacy) and CIC-IDS2017 (modern).
- To evaluate the ability of the models to predict in the situations of zero-day attack.
- To measure the zero-day detection performance in a leave-one-attack-out simulation framework that can be reproducible.

This study aims to offer a more feasible judgment of the effectiveness of AI-based IDS through the incorporation of dual-dataset assessment and standardized zero-day testing.

1.5 Contributions

The study has a number of valuable contributions to AI-driven intrusion detection. First, it outlines a comparative evaluation of intrusion detection models in the form of two-dataset comparisons in both legacy (NSL-KDD) and modern (CIC-IDS2017) traffic conditions, which allows a systematic study of the effect of the evolution of datasets on model behavior. It presents a reproducible zero-day simulation model using a leave-one-attack-out strategy, which offers a methodical method of measuring unseen attack detection, as opposed to using supervised accuracy measures only. The research compares several machine learning and deep learning models such as the Random Forest, XGBoost, Support Vector machine, and LSTM using compatible metrics of evaluation to create a fair and consistent comparison. It performs cross-dataset performance to determine the generalization ability of these models outside of a single benchmark environment. It provides practical information regarding the development of intrusion detection system performance and states the weaknesses of the existing AI-based methods in front of a zero-day threat. This work addresses a gap between high supervised accuracy and practical zero-day robustness, which contributes to the comprehensive understanding of the changing efficacy and real-life preparedness of AI-based intrusion detection systems.

2. Methodology

This paper follows a well-defined and repeatable experimentation model to measure the intelligence and performance of artificial intelligence-based intrusion detection systems in terms of supervised and zero-day environments. The approach combines two-data set assessment, pre-processing system, benchmark model

testing and controlled zero-day simulation plan. This framework proves to be a solution to the flaws of previous intrusion detection studies, where research commonly bases an evaluation on a single benchmark and heterogeneous testing procedures (Dai et al., 2024; Guo, 2022)

2.1 Datasets

Two commonly known intrusion dataset were chosen to be able to represent various generations of network traffic: NSL-KDD and CIC-IDS2017. The datasets are quite dissimilar in features, traffic realism, and attack diversity, so the model robustness can be studied within the changing cybersecurity contexts.

2.1.1 NSL-KDD

NSL-KDD is an improved set of KDD-99 dataset to minimize redundancy and class imbalance and still maintain its benchmarking value. It has 41 original features that detail the connection in the network such as the type of protocol, service, flag status, and statistical data like the number of connections and error rates. Some of these characteristics are categorical; hence, one-hot encoding was used, and the space of features extended to 122 numerical attributes.

In this study, NSL-KDD was set to binary classification with normal traffic as class 0 and all the types of attacks as class 1. Despite the fact that NSL-KDD is still actively utilized in the field of intrusion detection, the existing body of literature has pointed out that the structured and simulated nature of the NSLKDD can generate exaggerated supervised performance indicators when used in real traffic-related settings (Guo, 2022; Mirza et al., 2025). It can still be considered as a necessary standard of assessing model behavior in controlled experimental conditions.

2.1.2 CIC-IDS2017

CIC-IDS2017 is a current flow-based intrusion data which is set up to mirror real world enterprise network conditions. As opposed to NSL-KDD, based on artificial connections characteristics, CIC-IDS2017 uses 77 numerical attributes of two-way traffic flows. These aspects characterize flow duration, packet length statistics, inter-arrival times, flag counts, and other traffic ratios and are a more detailed characterization of network behavior.

The data set contains the latest attack scenarios like distributed denial-of-service (DDoS), brute-force attacks, intrusion, and web-based intrusion. As a result of class imbalance, stratified sampling was used before training in order to balance the malicious and benign cases. Such a method avoids model bias in favor of the majority classes but maintains representative distributions of attacks. The use of realistic traffic datasets is in line with the recent studies that propose the value of more current benchmarks in assessing the robustness of AI-based intrusion detectors (Abhijit et al., 2025; Nandiraju et al., 2025; Rodríguez et al., 2022).

2.1.3 Dataset Comparison

Table 1 provides the most significant features of the two data sets in order to emphasize the differences between the two generations and the structural distinctions between the two sets.

Table 1: Comparison Between NSL-KDD and CIC-IDS2017

Property	**NSL-KDD**	**CIC-IDS2017**
Era	Legacy	Modern
Feature Type	Mixed categorical and numerical	Flow-based numerical statistics
Traffic Realism	Simulated environment	Realistic enterprise traffic
Feature Count	122 (after encoding)	77
Attack Diversity	Structured and limited	Contemporary and diverse

The comparison shows that NSL-KDD represents the previous simulated network scenarios with designed characteristics, whereas CIC-IDS2017 represents the current dynamics of traffic using statistical flows characteristics. This contrast allows considering the impact of dataset evolution on the model performance and its generalization ability.

2.2 Data Preprocessing

To achieve economic and reliable assessment of the models, the preprocessing operations were standardized to apply on both datasets.

In the case of NSL-KDD, one-hot encoding was applied on the categorical attributes such as the protocol type, service and flag. This transformation enabled it to be compatible

AI and Machine Learning Advances
Volume 2, Issue 1
ISSN: 3067-3216

with numerical learning algorithms and retained categorical differences.

In the case of CIC-IDS2017, data cleaning was needed to deal with infinite values and missing entries that were created during the flow extraction. Without bound values had been replaced and then eliminated in order to ensure numerical stability.

StandardScaler was used as a feature scaling method to make the features have mean of 0 and a unit variance. Scaling was also used because of the division of the data into training and test to avoid the leakage of the data. This standardization process will make sure that model learning is not dominated by attributes that have larger ranges of numbers.

The binary label transformation was used throughout the two datasets where normal or benign traffic belongs to class 0 and malicious traffic belongs to class 1. Such standardization will allow having a direct comparison between supervised and zero-day evaluation scenarios.

2.3 Experimental Setup

The experimental setup consists of controlled classification and zero-day simulation to determine the capability of generalization.

Three traditional machine learning models were chosen: Random Forest, XGBoost and Support Vector Machine. Random Forest has been selected because it is an ensemble based on strength, and overfitting. XGBoost was chosen because of its gradient boosting optimization and good predictive abilities on tabular data. Support Vector Machine was added as a margin based classifier that can deal with high dimensional feature space. These are popular models used in intrusion detection studies and offer a moderate comparison of the ensemble and margin-based learning approaches (Dai et al., 2024).

Besides the conventional machine learning models, a Long Short-Term Memory (LSTM) network was introduced as the baseline of the deep learning. The architecture of LSTMs is designed to learn the complicated interactions between features and temporal relationships, which is why it is applicable in network traffic. The usage of LSTM can be attributed to the growing use of the deep learning methods in the study of zero-day detection (Nebhnani & Agrawal, 2025; Rodríguez et al., 2022).

A Leave-One-Attack-Out (LOAO) approach was utilized in order to replicate realistic zero-day conditions. Under this model, a single group of attack is omitted with the training data and is only presented in the testing process. The model is thus tested on its capability to identify new patterns of attacks that have never been seen before. Zero-Day Recall is calculated as the ratio of the correctly identified samples of unknown attacks to the total samples of unknown attacks. The approach offers a predictable and repeatable simulation of real-world zero-day detection conditions and overcomes the weaknesses witnessed in previous assessment models (Abhijit et al., 2025; Sarhan et al., 2021).

2.4 Evaluation Metrics

The model performance was evaluated based on various complementary measures, among which there are Accuracy, Recall, F1-score, ROC-AUC, False Positive Rate (FPR), and Zero-Day Recall. Although supervised measures give an idea of the overall detection performance of known attacks, the key performance measure put into focus is Zero-Day Detection Rate. This emphasis makes sure that they extend evaluation beyond inflated supervised accuracy up to realistic evaluation of the capability to detect unseen attacks. Through integrated and unified evaluation of supervised and zero-day evaluation, the methodology will provide a cohesive analysis of how AI-led intrusion detection systems change their effectiveness.

3. Results

The section reports the experimental results of supervised classification and zero-day simulation of the NSL-KDD and CIC-IDS2017 datasets. Findings are structured to include the performance under supervision, inter-dataset testing, and zero-day testing. All the figures used are directly related to the produced confusion matrices and ROC curves.

3.1 Supervised Performance on NSL-KDD

The controlled test of NSL-KDD shows almost perfect performance of all the tested models in terms of classification. Table 2 reflects the most important measures.

The findings show that Recall and F1-scores of Random Forest and XGBoost are very high and ROC-AUC scores of Random Forest and XGBoost are near to 1.0. The Support Vector machine is also high-performing with slightly lower recall and F1 than tree-based models.

The confusion matrix shows (Figure 1) that there is almost perfect separation between normal and attack traffic. There are minimal false positives and false

Table 2: Performance Metrics on NSL-KDD

Model	Recall	F1	ROC-AUC
RF	0.9986	0.9990	0.99999
XGB	0.9991	0.9993	0.99999
SVM	0.9929	0.9908	0.9989

negatives, which means that it has a high discriminative ability. Random Forest can also be used to successfully extract the feature interactions in the structured NSL-KDD data, indicating the separability of the dataset and the low complexity in supervised learning environment.

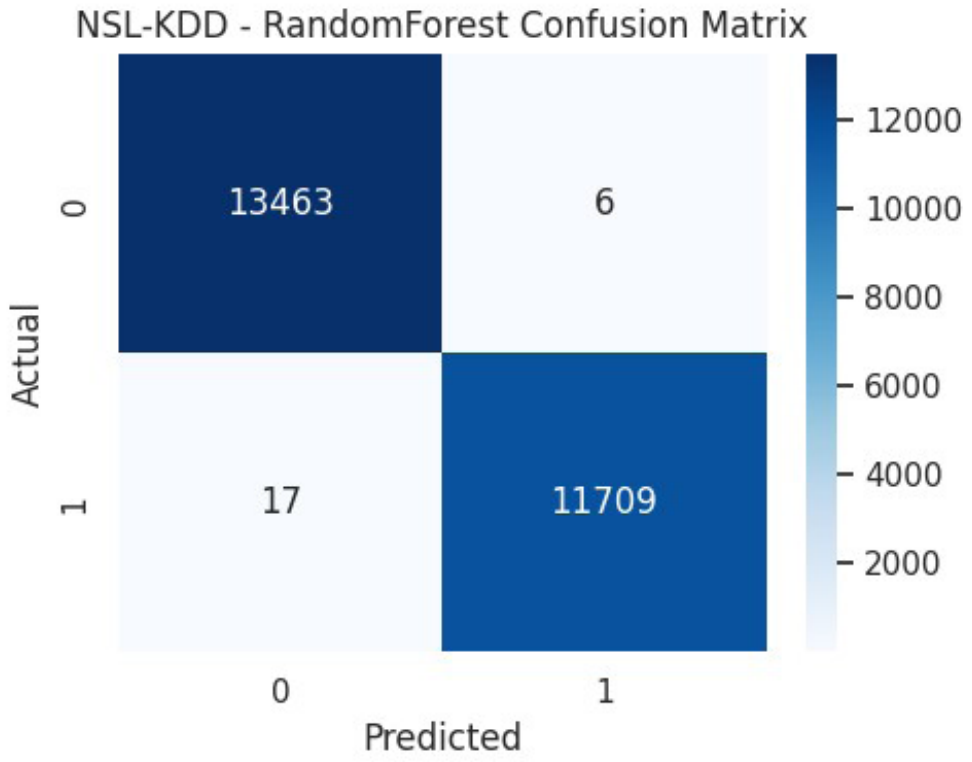

Figure 1: NSL-KDD Random Forest Confusion Matrix

XGBoost is close to perfect in the classification with very few cases of misclassification (Figure 2). The mechanism of boost increases sensitivity towards patterns of minorities, decreasing false negatives even more. This model has an engineered feature design of NSL-KDD which has a slightly better recall in comparison with other classifiers.

The SVM confusion matrix indicates that there are a little more false negatives as compared to ensemble techniques (Figure 3). Although it still provides great overall accuracy, margin-based classification seems not to be as responsive to fine details in attack patterns as tree-based ensembles are, and this is why it has relatively lower recall.

All the models have their ROC curves around the upper-left corner, which means the high use of discrimination. The curve validates that NSL-KDD can be very well separated when under a supervised environment. The approximate unity AUC scores indicate that the data might not be challenging enough to the contemporary machine learning algorithms (Figure 4).

In general, NSL-KDD is highly separable, false positive is minimal, and has properties of structured legacy benchmark data.

AI and Machine Learning Advances
Volume 2, Issue 1
ISSN: 3067-3216

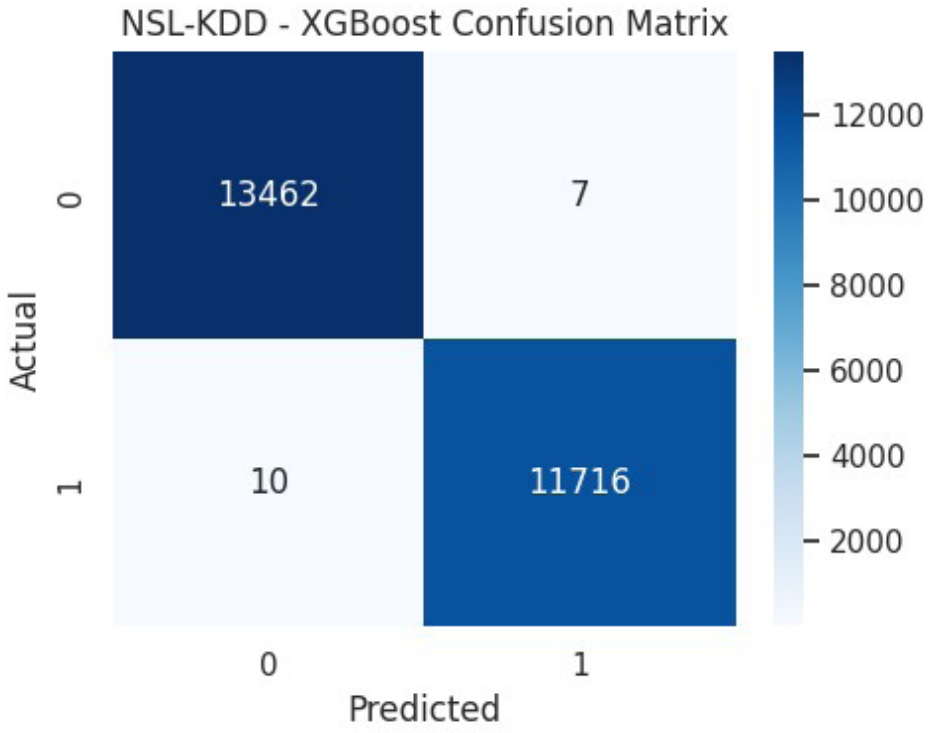

Figure 2: NSL-KDD XGBoost Confusion Matrix

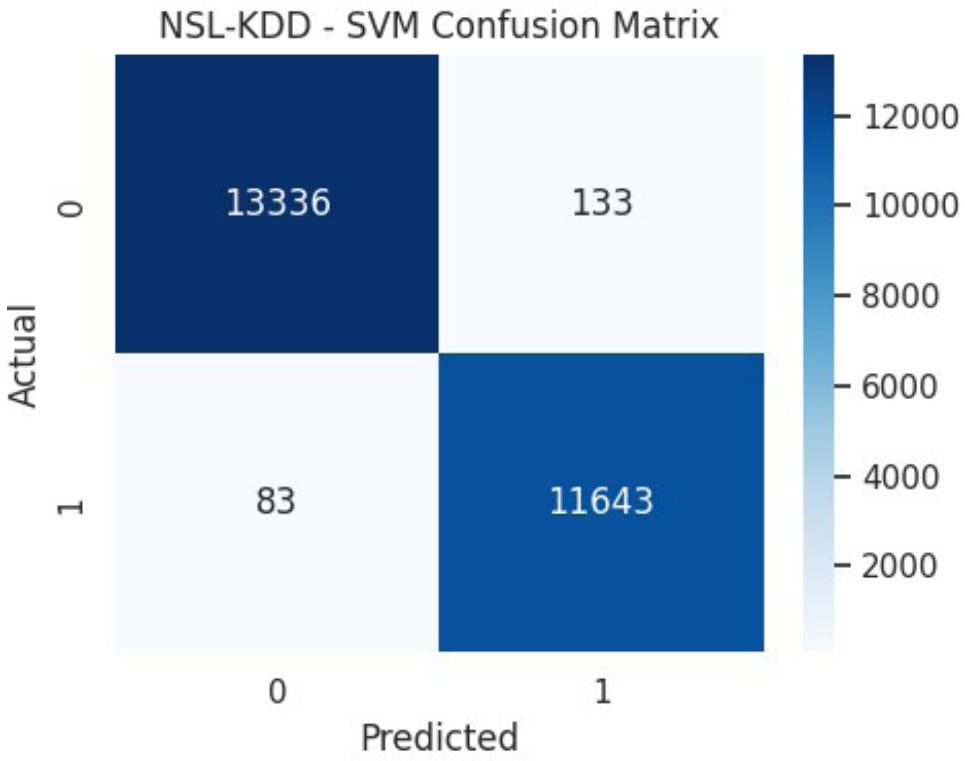

Figure 3: NSL-KDD SVM Confusion Matrix

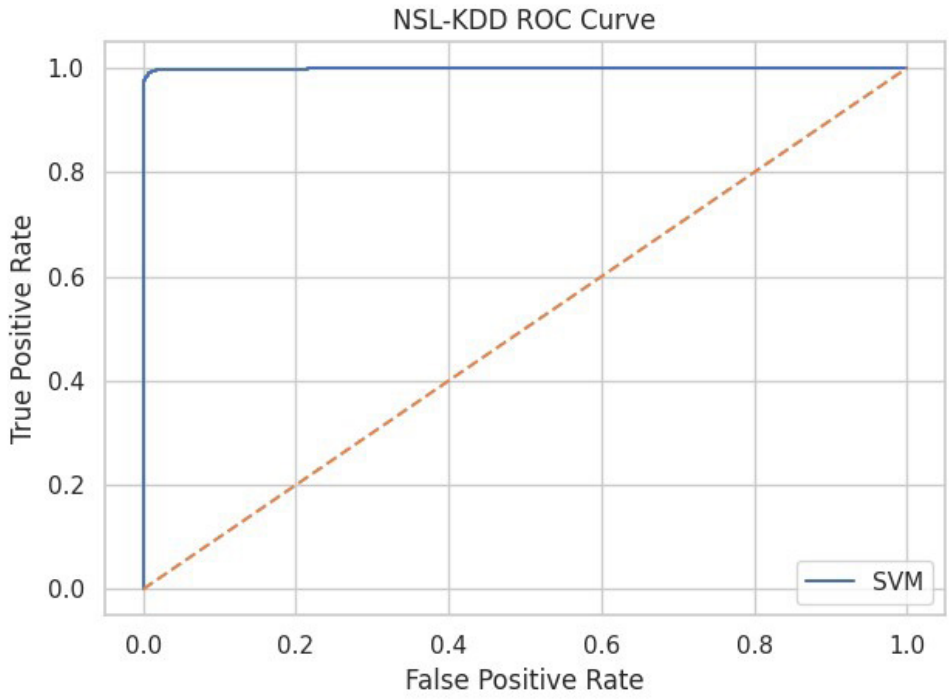

Figure 4: NSL-KDD ROC Curve

3.2 Supervised Performance on CIC-IDS2017

The obtained results on CIC-IDS2017 under the supervision are also strong but with a slightly higher number of misclassifications than with NSL-KDD because the dataset is more complex (Table 3).

Table 3: Performance Metrics on CIC-IDS2017

Model	Recall	F1	ROC-AUC
RF	0.9982	0.9982	0.99958
XGB	0.9992	0.9989	0.99991
SVM	0.9893	0.9817	0.9909

XGBoost had the best recall and AUC, and closely behind, Random Forest. The SNL-KDD performances have shown less performance degradation than the SVM performances indicating that SVM is more sensitive to the complexity of the dataset.

According to the confusion matrix, it shows a few falser positive and false negative than NSL-KDD (Figure 5). Even though the performance is good, the realistic enterprise traffic patterns add more classification challenges. Random Forest is resistant to feature diversity and flow-based complexity.

XGBoost has high recall rates having minimal misclassifications (Figure 6). The boosting mechanism successfully addresses non linear relationships of flow based features. The model is resilient to realistic traffic variability with the best overall performance compared to the considered classifiers.

SVM has a significant growth on misclassified examples than ensemble techniques. The decision boundary is a margin-based one, which seems to be less adaptable in the complex traffic distribution (Figure 7). This decrease in performance demonstrates the difficulty of realistic, high-dimensional flow statistics.

The ROC curves are very near to the ideal threshold and they exhibit a slight lower separation as compared to that of NSL-KDD. Decrease in AUC marginally is an

AI and Machine Learning Advances
Volume 2, Issue 1
ISSN: 3067-3216

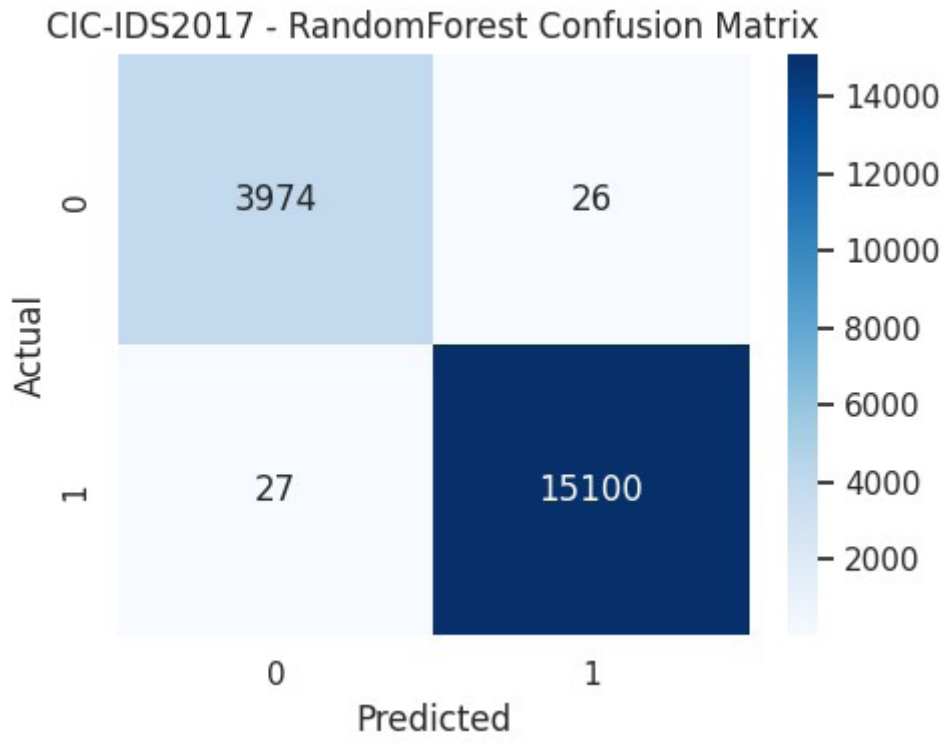

Figure 5: CIC-IDS2017 Random Forest Confusion Matrix

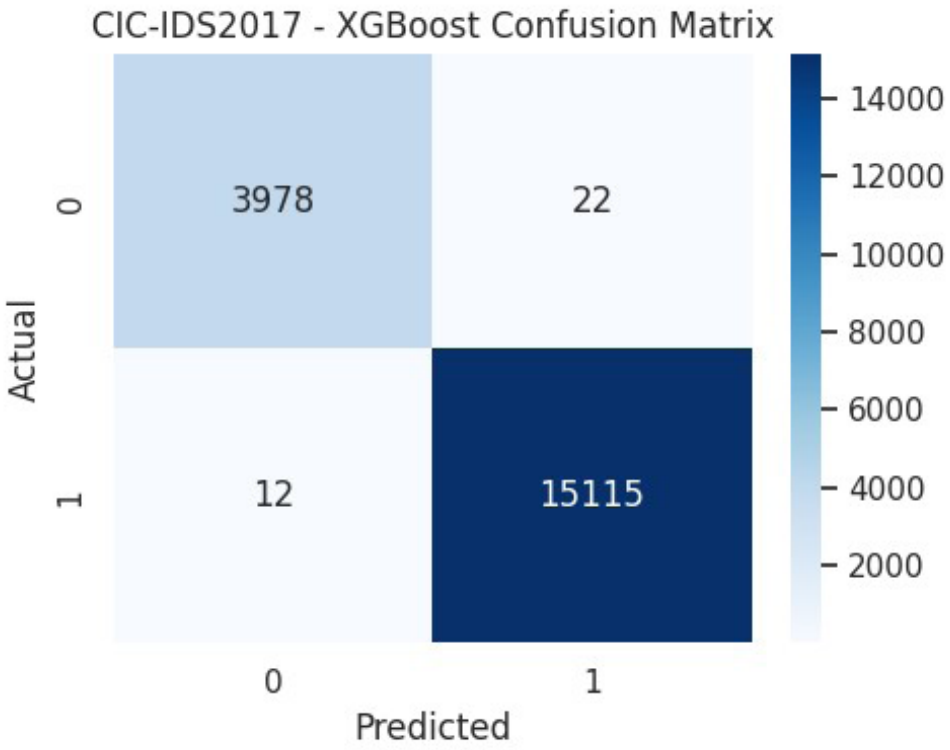

Figure 6: CIC-IDS2017 XGBoost Confusion Matrix

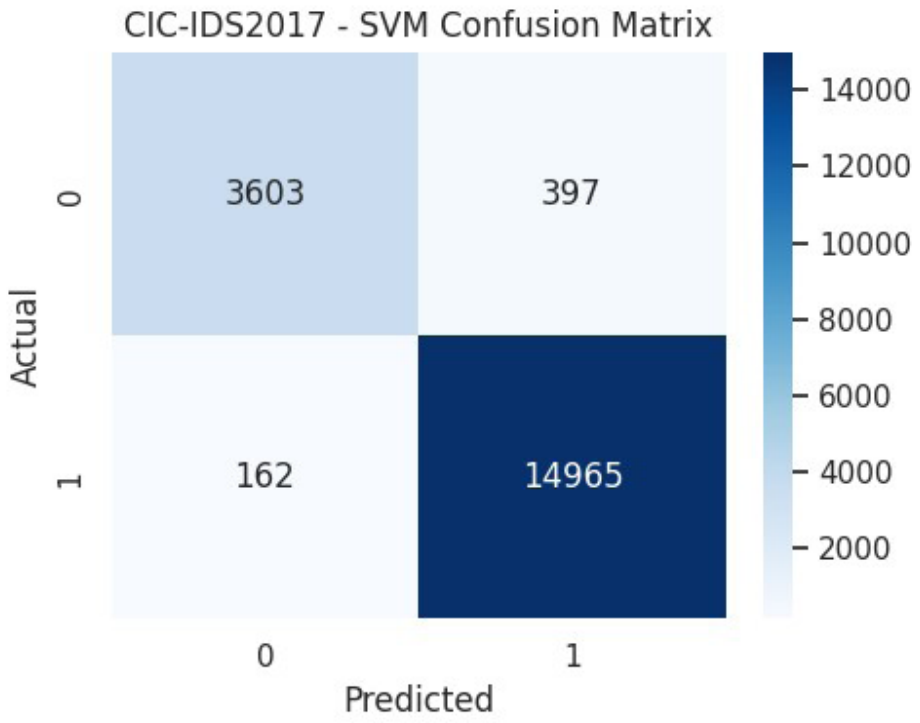

Figure 7: CIC-IDS2017 SVM Confusion Matrix

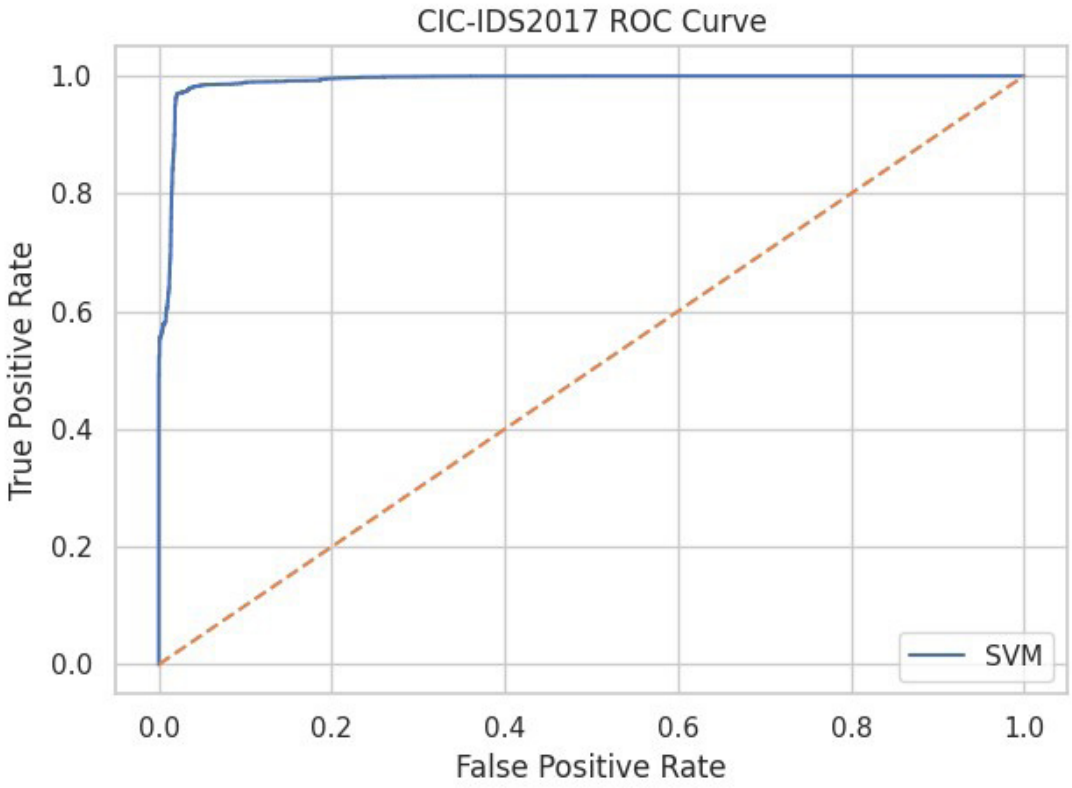

Figure 8: CIC-IDS2017 ROC Curve

AI and Machine Learning Advances
Volume 2, Issue 1
ISSN: 3067-3216

indicator that there is greater complexity in datasets. The current traffic conditions present more of a demanding classification environment to machine learning models. CIC-IDS2017 is more complex and realistic in nature and hence supervised performance is slightly lower (Figure 8).

3.3 Cross-Dataset Comparison

Cross-dataset comparisons were carried out in order to investigate the consistency of performance.

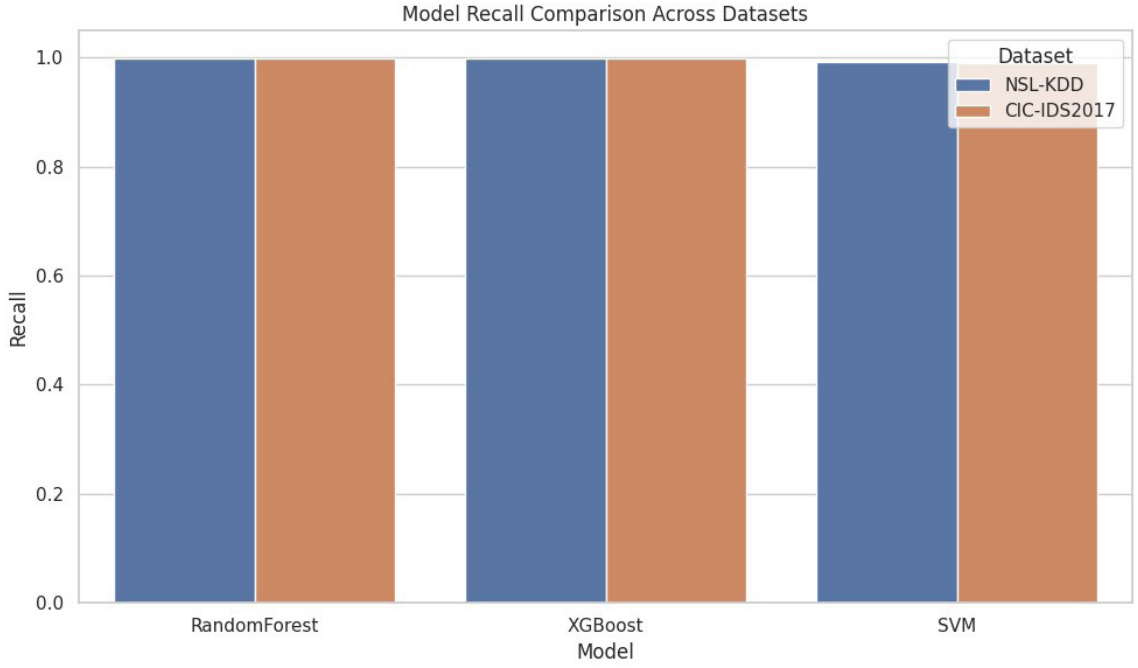

Figure 9: Recall Comparison Across NSL-KDD and CIC-IDS2017

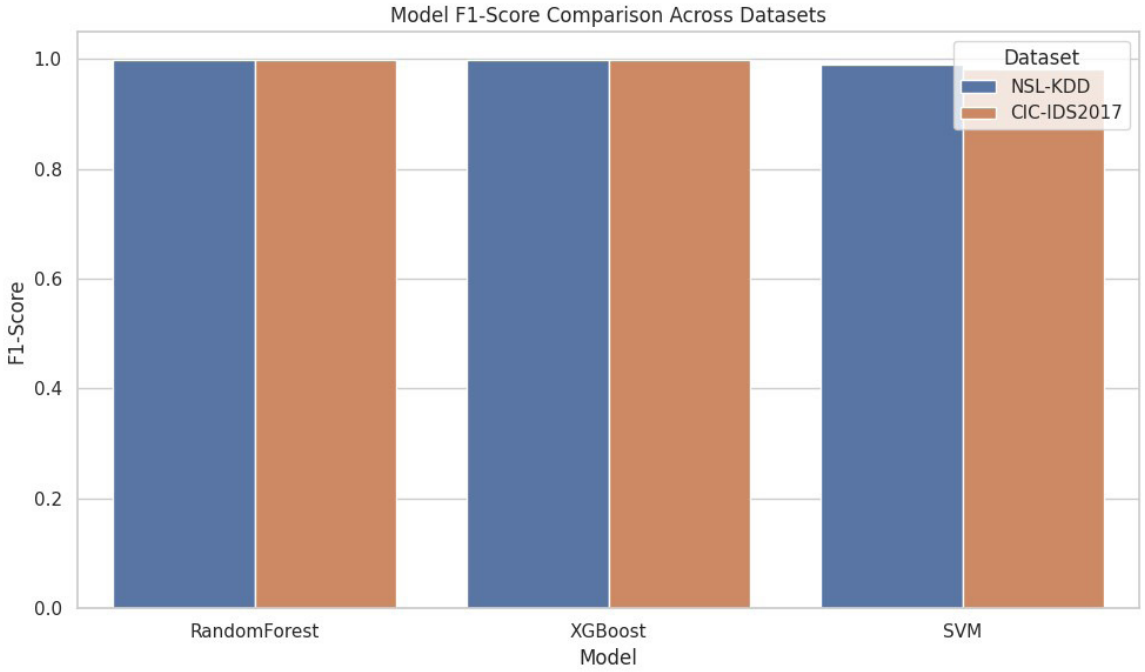

Figure 10: F1-Score Comparison Across NSL-KDD and CIC-IDS2017

There exists a high recall of tree based-models in both datasets, compared to high dataset sensitivity of SVM. The minor decrease in the recall of CIC-IDS2017 suggests that present-day traffic conditions introduce classification problems that are not present in old sets of data (Figure 9).

The trends in F1-scores assert homogeneous ensemble performance and comparative decay of SVM on the contemporary data. The findings indicate that ensemble-based learning is better generalized in changing traffic features, which adds more weight to their relevance in intrusion detection problems (Figure 10). The comparison across datasets demonstrates that the development of datasets affects the stability of the model, especially in case of margin-based classifiers.

3.4 Zero-Day Detection Results

Although the performance under the supervision has been exceptionally high, zero-day simulation demonstrates the performance degradation that can be measured (Table 4).

Table 4: Zero-Day Recall Across Models

Model	NSL ZD Recall	CIC ZD Recall
RF	0.842	0.731
XGB	0.864	0.748
LSTM	0.789	0.692

Zero-day analysis indicates that the recall is significantly lower than supervised ones, which proves that there is poor generalization to unknown forms of attack. Ensemble models based on trees are more robust than LSTM in conditions of zero-day. The decrease in the performance is more severe in CIC-IDS2017, which stresses the augmented complexity of the modern network traffic surroundings.

Zero-day recall is less than supervised recall in both datasets, which means that it does not generalize to unknown types of attacks. This decrease is more significant in CIC-IDS2017, as it has a higher diversity of attacks and a natural traffic handling pattern.

These results support the fact that high accuracy in the supervised mode does not directly relate to strong zero-day detection. The strength of the models is not the same, and ensemble methods are more effective than deep learning in unseen attack problems. The performance discrepancy shown highlights the necessity of assessment models that are based on zero-day resilience rather than supervised metrics.

4. Discussion

4.1 Evolution from Legacy to Modern IDS

The comparative analysis of NSL-KDD and CIC-IDS2017 reveals the developmental trend of intrusion detection data and its impact on the quality of the model. The observed near-perfect supervised performance on NSL-KDD validates long existing worries that old data sets can be too organized and easily disaggregated by current machine learning techniques (Guo, 2022; Üstebay, 2025). The artificial characteristics and the artificial traffic dynamics of NSL-KDD generate a classification setting where ensemble-based models have nearly perfect discrimination. Although this illustrates algorithmic functionality, it can be an over estimation of real-life strength.

By contrast, CIC-IDS2017 is based on more realistic network behavior in an enterprise, with recent attack vectors and statistical representations in form of flows. The marginal yet steady drop in CIC-IDS2017 supervised measures is an indication that the complexity of modern traffic presents more challenges in terms of classification. This fact was observed in accordance with the recent research that newer datasets are more likely to reflect the changing attack surfaces and behavioral variety (Ali et al., 2022; Rodríguez et al., 2022). The fact that connection level engineered capabilities are being replaced by flow-based statistical descriptions serves to emphasize the growing importance of flexible and powerful capabilities of detection. In general, the findings show that dataset realism plays a great role in determining the results of IDS evaluation.

4.2 AI vs Traditional ML

The results of the experiment indicate that tree-based ensemble models, namely, Random Forest and XGBoost tend to be more successful or equal to deep learning models in both data sets. These models obtained high recall and F1-scores and had high ROC-AUC.

AI and Machine Learning Advances
Volume 2, Issue 1
ISSN: 3067-3216

Ensemble approaches are also noise-resistant and can detect nonlinear interactions between features, which is especially useful with structured tabular network data (Alansary et al., 2025; Hairab et al., 2023; Wahed, 2025).

Though the LSTM model has shown competitive performance supervised, it was more sensitive towards the zero-day conditions. Deep learning systems are prone to generalizing better on larger and more diverse datasets and deteriorate when faced with distribution shifts (Hairab et al., 2023). Tree-based models do not seem susceptible to overfitting in the tabular intrusion detection setting. These results indicate that deep learning is promising, but traditional ensemble methods are also very competitive and more stable in the deployment of an IDS, in certain cases.

4.3 Zero-Day Detection Insights

The least important feature of this study is the simulation outcomes of zero-day. Even with almost perfect supervised metrics, every model had been observed to suffer significant recall degradation when tested on unseen classes of attacks. This proves that high accuracy under supervision does not mean that they are robust to new threats. The resulting decrease in performance, especially in the more difficult CIC-IDS2017 dataset, points to the difficulty of generalization when operating in changing cybersecurity environments.

The issue of feature representation is a key part of zero-day detection. The discriminative patterns implicitly encoded by engineered attributes of NSL-KDD, can be used to directly separate anomalies, which is not the case with the statistical flow features of CIC-IDS2017, which uses models to estimate behavioral deviations of more global traffic distributions. As has been mentioned in previous studies, the ability of a model to identify new threats depends heavily on the diversity of the data used (Alam & Ahmed, 2023; Hashim et al., 2025). The gap in terms of zero-day recalls that is observed in this work supports the significance of the evaluation frameworks that go beyond the supervised classification to the realistic robustness assessment.

4.4 Practical Implications

Operational wise, the results provide valuable insights in deployment. The tree-based models were found to be especially high-performing and had relatively low methods of computation, and thus, could be used in detecting intrusion in real-time systems. XGBoost, most notably, was found to have high recall and inference latency that can be easily managed, implying the possibility of being practically feasible in the field of enterprise applications.

LSTM and other deep learning models consume more resources and time to train. They might be able to capture complex dependencies, but marginal gains in supervised performance do not always explain higher resource usage in resource-constrained settings. Moreover, the presence of small false positive rates may have a profound effect on the functioning systems, which causes the fatigue of alerts and the decrease of analyst confidence. Hence, it is crucial to balance detection accuracy and false positive control so that IDS can be practically implemented.

4.5 Limitations

There are a number of constraints that need to be realized. To begin with, NSL-KDD is an older dataset that does not necessarily represent the behavior of networks today, which can exaggerate supervised performance outcomes (Almuflih et al., 2024; Hairab et al., 2023; Moreno et al., 2024). Stratified sampling used on CIC-IDS2017, though required due to the balance of the classes, could cause a change in the initial distribution of the traffic. Third, binary classification simplifies the taxonomy of attacks, which may obscure subtle behavior in the model in a particular set of attacks. Lastly, the experimental configuration is a controlled offline test, in the real world deployment scenarios, traffic patterns and adversarial adaptations are dynamic and are not captured in this model.

These limitations notwithstanding, the dual-dataset assessment and a zero-day simulation could still offer an informative idea of the changing effectiveness of AI-based intrusion detection systems.

5. Conclusion

This experiment compared AI evolution and effectiveness of intrusion detection systems by using a two-dataset experimental design using a traditional benchmark (NSL-KDD) and a more recent flow-based data (CIC-IDS2017). The monitored performance of the assessment showed that modern machine learning and deep learning models have significantly high performance on ordered benchmark datasets. Forest machine models, especially the Random Forest and XGBoost, were almost perfect in recall and ROC-AUC, indicating that they have high discriminatory ability in controlled classification.

The experiments with the zero-day simulation have shown

that there is a severe limitation: the large supervised accuracy does not always imply the strong recognition of the still unknown types of attacks. In both datasets, every assessed model was subject to an observable decrease in performance in the leave-one-attack-out model. It was larger in the modern CIC-IDS2017 data, which emphasizes the greater complexity and changing nature of realistic enterprise network traffic. These results highlight the fact that generalization is one of the critical issues in AI-assisted intrusion detection.

The comparative study has also shown that the characteristics of the datasets have a considerable impact on the behavior of the model. Older datasets can be overshadowing the real-world preparedness whereas new datasets demonstrate weaknesses in robustness and adaptability. This study better informs of the behavior of intrusion detection performance development with respect to generational benchmarks through the incorporation of cross-dataset assessment, in combination with a reproducible zero-day simulation plan.

Although AI-based models of IDS demonstrate a high level of supervised performance, the challenge of enhancing the resilience of zero-days is critical to the practical implementation. The research of the future must focus on different data sets, adaptive learning methods, and assessment models that will focus on generalization in the changing threat landscapes.

6. Declarations

Ethics approval and consent to participate: Not applicable. This study involves the computational analysis of publicly available, de-identified benchmark datasets (NSL-KDD and CIC-IDS2017) and does not involve human participants, animals, or sensitive personal data.

Consent for publication: Not applicable. The manuscript does not contain any individual person's data in any form (including individual details, images, or videos).

Availability of data and material: The datasets analysed during the current study are available in the following repositories:

1. The NSL-KDD dataset: https://www.unb.ca/cic/datasets/nsl.html
2. The CIC-IDS2017 dataset: https://www.unb.ca/cic/datasets/ids-2017.html. All algorithms and models used are implemented via standard Python libraries (Scikit-learn, XGBoost) as described in the methodology.

Conflicts of Interest: The authors declare that they have no competing interests.

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Authors' contributions: KRISHNAM NIMMALA was responsible for the conceptualisation of the study, experimental design, data preprocessing, implementation of machine learning models (Random Forest, XGBoost, SVM), formal analysis of zero-day simulations, and the writing of the original draft. The author has read and approved the final manuscript.

Acknowledgements: The authors would like to acknowledge the Canadian Institute for Cybersecurity (CIC) for providing the benchmark datasets used in this research.

References

Abhijit, C. S., Jerusha, A., Ibrahim, S. S., & Varadharajan, V. (2025). Federated transfer learning for rare attack class detection in network intrusion detection systems. Scientific Reports, 15. https://doi.org/10.1038/s41598-025-02068-x

Alam, N., & Ahmed, M. (2023). Zero-day Network Intrusion Detection using Machine Learning Approach. International Journal on Recent and Innovation Trends in Computing and Communication. https://doi.org/10.17762/ijritcc.v11i8s.7190

Alansary, S. A., Ayyad, S., Talaat, F., & Saafan, M. (2025). Emerging AI threats in cybercrime: a review of zero-day attacks via machine, deep, and federated learning. Knowledge and Information Systems, 67, 10951-10987. https://doi.org/10.1007/s10115-025-02556-6

Ali, S., Rehman, S., Imran, A., Adeem, G., Iqbal, Z., & Kim, K.-I. (2022). Comparative Evaluation of AI-Based Techniques for Zero-Day Attacks Detection. Electronics. https://doi.org/10.3390/electronics11233934

Almuflih, A., Abdullayev, I., Bakhvalov, S., Shichiyakh, R., Dash, B., Rao, K., & Bansal, K. (2024). Securing IoT devices with zero day intrusion detection system using binary snake optimization and attention based bidirectional gated recurrent classifier. Scientific

Reports, 14. https://doi.org/10.1038/s41598-024-80255-y

Dai, Z., Por, L., Chen, Y.-L., Yang, J., Ku, C. S., Alizadehsani, R., & Pławiak, P. (2024). An intrusion detection model to detect zero-day attacks in unseen data using machine learning. PLOS One, 19. https://doi.org/10.1371/journal.pone.0308469

Guo, Y. (2022). A review of Machine Learning-based zero-day attack detection: Challenges and future directions. Computer communications, 198. https://doi.org/10.1016/j.comcom.2022.11.001

Hairab, B. I., Aslan, H., Elsayed, M. S., Jurcut, A., & Azer, M. (2023). Anomaly Detection of Zero-Day Attacks Based on CNN and Regularization Techniques. Electronics. https://doi.org/10.3390/electronics12030573

Hashim, K. A., Yussoff, Y. B. M., & Shahbudin, S. B. (2025). Mitigating Zero-Day Vulnerabilities in IIoT Systems: Challenges and Advances in AI-Powered Intrusion Detection Systems. Mesopotamian Journal of CyberSecurity. https://doi.org/10.58496//mjcs/2025/063

Kansal, S. (2025). Utilizing Deep Learning Techniques for Effective Zero-Day Attack Detection. Economic Sciences. https://doi.org/10.69889/m3jzbt24

Karn, A. L., Ghanimi, H., Iyengar, V., Siddiqui, M. S., Alharbi, M., Alroobaea, R., Yousef, A., & Sengan, S. (2025). Applying the defense model to strengthen information security with artificial intelligence in computer networks of the financial services sector. Scientific Reports, 15. https://doi.org/10.1038/s41598-025-15034-4

Lin, X. (2025). A Survey of AI-Based Zero-Day Attack Detection Methods. Applied and Computational Engineering. https://doi.org/10.54254/2755-2721/2025.po25664

Mirza, A., Arshad, S., Yousaf, M., & Azam, M. A. (2025). ZDBERTa: Advancing Zero-Day Cyberattack Detection in Internet of Vehicle with Zero-Shot Learning. Comput., 14, 424. https://doi.org/10.3390/computers14100424

Moreno, J. F. C., Rizzardi, A., Sicari, S., & Coen-Porisini, A. (2024). NERO: NEural algorithmic reasoning for zeRO-day attack detection in the IoT: A hybrid approach. Comput. Secur., 142, 103898. https://doi.org/10.1016/j.cose.2024.103898

Nandiraju, S. K. K., Chundru, S. K., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., & Kakani, A. B. (2025). Enhancing Cybersecurity: Zero-Day Attack Detection in Network Traffic with Deep Learning Model. Asian Journal of Research in Computer Science. https://doi.org/10.9734/ajrcos/2025/v18i7734

Nebhnani, N., & Agrawal, S. (2025). ProEn-XAI: A High-Precision IDS Model for Zero-Day Attack Detection Using Hybrid Deep Learning and SHAP-LIME Interpretability. International Journal of Environmental Sciences. https://doi.org/10.64252/cpzyn974

Rodríguez, E., Valls, P., Otero, B., Costa, J. J., Verdú, J., Pajuelo, M. A., & Canal, R. (2022). Transfer-Learning-Based Intrusion Detection Framework in IoT Networks. Sensors (Basel, Switzerland), 22. https://doi.org/10.3390/s22155621

Sarhan, M., Layeghy, S., Gallagher, M., & Portmann, M. (2021). From zero-shot machine learning to zero-day attack detection. International Journal of Information Security, 22, 947-959. https://doi.org/10.1007/s10207-023-00676-0

Üstebay, S. (2025). Enhancing Zero-Day Attack Detection in IoT Networks via Isolation Forest and Ensemble Tree Models. ELECTRICA. https://doi.org/10.5152/electrica.2025.24177

Wahed, M. A. (2025). AI-Enhanced Threat Intelligence for Proactive Zero-Day Attack Detection. Gamification and Augmented Reality. https://doi.org/10.56294/gr2025112