SSR Journal of Artificial Intelligence (SSRJAI)

OPEN ACCESS

Homepage: https://ssrpublisher.com/ssrjai/

Volume 3, Issue 8, 2026 ISSN: 3049-0413

Adversarially Robust Intrusion Detection in Software-Defined Networks Using a Resilient Adversarial Deep Neural Network with Stochastic Gradient Langevin Dynamics

Aisha Lawal Yusuf[1], Uche M. Mbanaso[2], Makinde Julius[3], Garba S. Abdullahi[4]

[1]Department of Computer Science, Faculty of Computing, Nile University of Nigeria, Abuja, Nigeria

[2]Department of Computer Science, Nasarawa State University, Keffi, Nigeria

[3]Department of Computer Science, Baze University, Abuja, Nigeria

[4]Centre for Cyberspace Studies, Nasarawa State University Keffi, Nigeria

Received: *05.07.2026* | ***Accepted:*** *10.08.2026* | ***Published:*** *13.08.2026*

***Corresponding author:** Aisha Lawal Yusuf

DOI: *10.5281/zenodo.21916729*

Abstract **Original Research**

Software-Defined Networking (SDN) has transformed network management by separating the control plane from the data plane, thereby improving flexibility, programmability, and centralized control. However, the centralized architecture also introduces security challenges, making SDN environments attractive targets for cyberattacks. Although deep learning-based Intrusion Detection Systems (IDSs) have demonstrated high detection performance, they remain vulnerable to adversarial attacks that can drastically degrade classification accuracy. This study presents an adversarially robust intrusion detection system for SDN by incorporating a Resilient Adversarial Deep Neural Network (RANetDNN) with Stochastic Gradient Langevin Dynamics (SGLD) optimization approach. The model was implemented and evaluated using the InSDN dataset and compared with a Deep Neural Network (DNN) baseline. Performance was evaluated using Accuracy, Precision, Recall, F1-score, ROC-AUC and adversarial robustness under Fast Gradient Sign Method (FGSM) attacks. Experimental results show that while the baseline DNN experienced a drastic reduction in F1-score under increasing adversarial perturbations, the RANetDNN-SGLD model consistently maintained high detection performance and robustness across all evaluated perturbation levels. The experimental results demonstrate that the RANetDNN-SGLD model achieved excellent intrusion detection performance while showing significantly greater robustness against FGSM adversarial attacks than the baseline Deep Neural Network. Comparative evaluation showed that this approach maintained consistently high F1-score values under increasing adversarial perturbations, highlighting its effectiveness for securing Software-Defined Network environments against both conventional cyber threats and adversarial attacks.

Keywords: Software-Defined Networking (SDN), Intrusion Detection System (IDS), Resilient Adversarial Network (RANet), Stochastic Gradient Langevin Dynamics (SGLD), Deep Learning, Deep Neural Network.

1.Introduction

Software-Defined Networking (SDN) is a transformative networking paradigm that separates the control plane from the data plane, enabling a centralized network management, programmability, flexibility, and efficient resource utilization. Unlike other traditional networks, where control functions are distributed across networking devices, SDN provides a logically centralized controller that makes

network configuration and management easier while supporting dynamic traffic changes and rapid deployment of new network services. These capabilities have enhanced the adoption of SDN in cloud computing, data centers, enterprise networks, Internet of Things (IoT) environments, and next-generation communication infrastructures. However, the centralized architecture that gives SDN its operational advantages also introduces significant security challenges, as the SDN controller becomes a high-value target for cyberattacks. Therefore, ensuring robust network security has become one of the most critical research issues in SDN environments (Jawad, 2018).

Intrusion Detection Systems play a vital role in protecting SDN infrastructures by monitoring network traffic and identifying malicious activities before they compromise network operations. Recent advances in Artificial Intelligence (AI), particularly Machine Learning and Deep Learning, have improved intrusion detection performance by learning complex traffic patterns and detecting unseen attacks. Deep learning models such as Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRUs) have demonstrated significant accuracy in SDN intrusion detection. Nevertheless, despite their high detection performance, these models remain vulnerable to adversarial attacks, where carefully crafted perturbations can manipulate model predictions and reduce detection accuracy (Mujahid et al., 2025).

Several recent studies have shown that deep learning techniques for SDN intrusion detection. (Hadi & Mohammed, 2022) proposed a Network Intrusion Detection System based on multiple deep learning classifiers and achieved high binary classification accuracy using the NSL-KDD dataset. (Chaganti et al., 2023) developed an LSTM-based intrusion detection framework for SDN-enabled Internet of Things networks and highlighted adversarial attacks as an important direction for future research. (Ataa et al., 2024) evaluated Transformer and CNN-LSTM architectures using the InSDN dataset and demonstrated that deep learning models can achieve excellent intrusion detection performance in SDN environments. Furthermore,(Doğan et al., 2025) conducted a comprehensive review of AI-driven SDN security and emphasized that improving adversarial robustness remains one of the major unresolved challenges. (Alyahyan, 2025) subsequently proposed the Resilient Adversarial Network (RANet), demonstrating that adversarial resilience training can significantly improve the robustness of deep learning models against adversarial perturbations.

Although these studies have significantly advanced SDN intrusion detection, most existing approaches primarily focus on maximizing classification accuracy under normal operating conditions without explicitly addressing adversarial robustness. Furthermore, the integration of Resilient Adversarial Networks (Saini & Chennamaneni, 2023) with Stochastic Gradient Langevin Dynamics (SGLD) (Li et al., 2024) optimization for SDN intrusion detection has not been adequately investigated. This limitation creates a significant research gap because intrusion detection models deployed in real-world SDN environments must remain reliable even when confronted with adversarially manipulated network traffic.

This study addresses this gap by proposing an adversarially robust intrusion detection system for Software-Defined Networks through the integration of a Resilient Adversarial Deep Neural Network (RANetDNN) with Stochastic Gradient Langevin Dynamics (SGLD) optimization. The proposed model is implemented and evaluated using the InSDN dataset (Elsayed & Jurcut, 2021) and compared with a conventional Deep Neural Network baseline under both normal and adversarial conditions using the Fast Gradient Sign Method (FGSM). The performance of the proposed model is assessed using **Accuracy, Precision, Recall, F1-score, ROC-AUC, and robustness against FGSM adversarial attacks** to determine its effectiveness in improving SDN intrusion detection performance.

2.Materials and Methods

2.1 Research Design

In order to develop an adversarially robust intrusion detection system for Software-Defined Networks, this study used an experimental research approach. To increase the resilience of intrusion detection against adversarial assaults, the system combines Stochastic Gradient Langevin Dynamics (SGLD) optimization with a Resilient Adversarial Deep Neural Network (RANetDNN). Model creation, adversarial sample generation using the Fast Gradient Sign Method (FGSM), dataset preparation, feature preprocessing, model training, and performance assessment were all part of the implementation. The InSDN dataset was used to test and train the model, which was developed with Python and the PyTorch deep learning framework. To offer a comparative evaluation of detection performance and adversarial robustness, a traditional Deep Neural Network (DNN) was used as the baseline model.

2.2 Dataset Description

The InSDN dataset, a publicly available benchmark dataset created especially for evaluating intrusion detection systems in Software-Defined Networking (SDN) studies, was used for the studies. The dataset is suitable for creating and assessing deep learning-based intrusion detection models as it includes actual inSDN network traffic produced under both typical and attack states. Network flow records that shows both normal traffic and various types of cyberattacks make up the InSDN dataset. Distributed Denial-of-Service (DDoS), Denial-of-Service (DoS), Probe, Brute Force Attack (BFA), Web Attack, Botnet, and User-to-Root (U2R) assaults are some of these types of attacks. A thorough assessment of intrusion detection algorithms under various network threat scenarios is made possible by the variety of attack types.

For this study, the dataset was obtained from three CSV files, namely Normal data.csv, OVS.csv, and metasploitable2.csv, which were merged to create a single dataset for analysis. During preprocessing, duplicate records, missing values, and irrelevant attributes were removed, and the attack labels were converted into a binary classification problem consisting of Normal and Attack classes. Feature standardization was performed to normalize the numerical attributes before model training. The processed dataset was divided into training and testing subsets to enable unbiased performance evaluation of both the baseline Deep Neural Network (DNN) and the proposed RANetDNN-SGLD model.

2.3 Data Pre-processing

To enhance data quality and security with the suggested deep learning models, the InSDN dataset underwent a number of preparation procedures before model training. Data cleaning, feature selection, label transformation, feature scaling, and dataset segmentation made up the preparation pipeline.

First, a single dataset was created by combining the three source datasets (Normal data.csv, OVS.csv, and metasploitable-2.csv). To increase dataset consistency, duplicate records, missing values, and unnecessary characteristics were eliminated. After that, the attack labels were converted into a binary classification issue by designating malicious traffic as 1 and regular network traffic as 0. This change kept all attack instances inside a single attack class while restructuring the intrusion detection process. To ensure that all numerical features contributed equally during model training, feature standardization was performed using the Standard Scaler technique. This process normalized each feature to have zero mean and unit variance, preventing features with larger numerical ranges from dominating the learning process.

To address class imbalance during model training, a weighted Binary Cross-Entropy Loss function was employed. The positive class weight was computed from the class distribution of the training dataset and incorporated into the loss function, thereby assigning greater importance to the minority class during optimization.

Following preprocessing, the dataset was randomly divided into training and testing subsets. The training dataset was used for model learning and parameter optimization, while the

testing dataset was reserved for evaluating the generalization performance of both the baseline Deep Neural Network (DNN) and the RANetDNN-SGLD model.

2.4 RANetDNN-SGLD Architecture

The intrusion detection model incorporates a Resilient Adversarial Deep Neural Network (RANetDNN) with Stochastic Gradient Langevin Dynamics (SGLD) optimization to improve the robustness of intrusion detection in Software-Defined Networking (SDN) environments. The model was designed to learn normal and malicious network traffic while maintaining resilience against adversarial perturbations.

The architecture consists of an input layer corresponding to the selected network traffic features, multiple fully connected hidden layers with Rectified Linear Unit (ReLU) activation functions, dropout regularization, and a single-neuron output layer for binary classification. The output layer produces logits, which are converted into probabilities using the sigmoid activation function during inference.

To improve robustness, adversarial training was incorporated using the Fast Gradient Sign Method (FGSM). During each training iteration, adversarial examples were generated by introducing perturbations in the direction of the gradient of the loss function. The model was subsequently trained using both clean and adversarial samples. The final training loss was computed as the average of the clean loss and adversarial loss, allowing the model to learn robust feature representations while maintaining classification performance on legitimate network traffic.

Model optimization was performed using Stochastic Gradient Langevin Dynamics (SGLD). Unlike stochastic gradient descent, SGLD injects controlled Gaussian noise into the parameter update process after an initial burn-in period. During the first six training epochs, optimization was performed without stochastic noise to facilitate stable convergence. Thereafter, Gaussian noise was introduced during the sampling phase, enabling broader exploration of the parameter space and reducing convergence to sharp minima that are often associated with poor generalization and increased vulnerability to adversarial attacks.

To further improve learning on the imbalanced dataset, a weighted Binary Cross-Entropy Loss function was employed. The positive class weight was computed directly from the class distribution of the training dataset and incorporated into the loss function, thereby reducing the bias toward the majority class during optimization.

The complete training process was implemented using the PyTorch deep learning framework, while model performance was evaluated using Accuracy, Precision, Recall, and F1-score.

2.5 Experimental Setup

The proposed RANetDNN-SGLD model was implemented using Python with the PyTorch deep learning framework. Model training was performed for 30 epochs using a batch size of 128 and an initial learning rate of 0.0001. Stochastic Gradient Langevin Dynamics (SGLD) was employed as the optimization algorithm to improve parameter exploration, model generalization, and uncertainty estimation.

The training process consisted of two phases. During the initial burn-in phase (first six epochs), the model was optimized without stochastic noise to ensure stable convergence. After the burn-in period, controlled Gaussian noise was introduced during parameter updates as part of the SGLD sampling phase to improve robustness and prevent convergence to sharp local minima.

To enhance resistance against adversarial attacks, the Fast Gradient Sign Method (FGSM) was incorporated during training. Adversarial samples were generated by perturbing the input data using a small perturbation factor, and the model was trained simultaneously on both clean and adversarial samples. The final training loss was computed by combining the losses from both sample types.

Model performance was evaluated on an independent test dataset using Accuracy,

Precision, Recall, F1-score. To obtain the most effective classification boundary, threshold optimization was performed on the validation dataset before the final evaluation on the test dataset.

3.Results and Discussion

3.1 Performance of the Baseline Deep Neural Network

To establish a benchmark for evaluating the effectiveness of the proposed RANetDNN-SGLD model, a conventional Deep Neural Network (DNN) was implemented and evaluated using the InSDN dataset. The baseline model was trained under the same experimental conditions as the proposed model to ensure a fair comparison. Performance was assessed using Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix analysis.

The experimental results presented in Table 1 indicate that the baseline DNN achieved excellent classification performance on the test dataset, recording an accuracy of 99.96%, precision of 99.98%, recall of 99.97%, and F1-score of 99.97%. The model also achieved a high ROC-AUC value, indicating excellent discrimination between normal and malicious network traffic. These findings demonstrate that the baseline model performs effectively under normal operating conditions provides a strong benchmark for evaluating the effectiveness of the RANetDNN-SGLD model."

Table 1 Performance of Baseline DNN Model on the InSDN Dataset

Metric	**Value** %
Accuracy	99.96%
Precision	99.96%
Recall	99.99%
F1-Score	99.97%
ROC-AUC	99.98%

Although the baseline DNN achieved excellent performance under normal operating conditions, these evaluation metrics alone do not provide sufficient evidence of model robustness. Deep learning models are known to be susceptible to adversarial perturbations that can significantly reduce detection performance even when classification accuracy on clean data is high. Consequently, additional evaluation under adversarial conditions is necessary to determine the reliability of the intrusion detection model in real-world SDN environments.

3.2 Performance of the Proposed RANetDNN-SGLD Model

The proposed RANetDNN-SGLD model was evaluated using the same experimental configuration as the baseline Deep Neural Network to ensure a fair performance comparison. The model integrates adversarial training using the Fast Gradient Sign Method (FGSM) with Stochastic Gradient Langevin Dynamics (SGLD) optimization to improve robustness while maintaining high intrusion detection performance.

The evaluation results presented in Table 2 demonstrate that the proposed model achieved excellent classification performance on the InSDN dataset. The high Accuracy, Precision, Recall, F1-score, and ROC-AUC values indicate that the proposed approach effectively distinguishes between normal and malicious network traffic.

The strong performance of the proposed model can be attributed to the integration of adversarial

training and SGLD optimization. Adversarial training enabled the model to learn discriminative features from both clean and perturbed samples, thereby improving resilience against adversarial manipulation. Additionally, the stochastic parameter updates introduced by SGLD enhanced model generalization by encouraging exploration of flatter regions of the optimization landscape, reducing the likelihood of overfitting and improving robustness.

Table 2 Performance of Proposed RANetDNN-SGLD Model

Metric	Value %
Accuracy	99.42%
Precision	99.42%
Recall	99.86%
F1-Score	99.64%
ROC-AUC	99.86%

3.3 Comparative Performance Analysis

To evaluate the effectiveness of the proposed approach, the performance of the RANetDNN-SGLD model was compared with that of the baseline Deep Neural Network (DNN) using the same training and testing datasets. The comparison was based on Accuracy, Precision, Recall, F1-score, and ROC-AUC, ensuring a comprehensive assessment of the classification performance of both models.

The comparative results presented in Table 3 indicate that both models achieved excellent intrusion detection performance under normal operating conditions. The baseline DNN demonstrated high classification accuracy, while the proposed RANetDNN-SGLD model achieved comparable or improved performance across the evaluated metrics. These findings indicate that incorporating adversarial training and SGLD optimization does not compromise the model's ability to accurately distinguish between normal and malicious network traffic.

Furthermore, the results suggest that the proposed model maintains strong detection capability while introducing additional robustness mechanisms that enhance resilience against adversarial attacks. This demonstrates that improving adversarial robustness can be achieved without sacrificing overall classification performance, making the proposed approach suitable for deployment in security-critical SDN environments.

The findings of this study are consistent with previous research demonstrating the effectiveness of deep learning techniques for intrusion detection in Software-Defined Networks

Table 3. Comparative Performance of the Baseline DNN and Proposed RANetDNN-SGLD Model

Metric	Baseline DNN (%)	RANetDNN-SGLD (%)
Accuracy	99.96	99.42

Precision	99.96	99.42
Recall	99.99	99.86
F1-Score	99.97	99.64
ROC-AUC	99.98	99.86

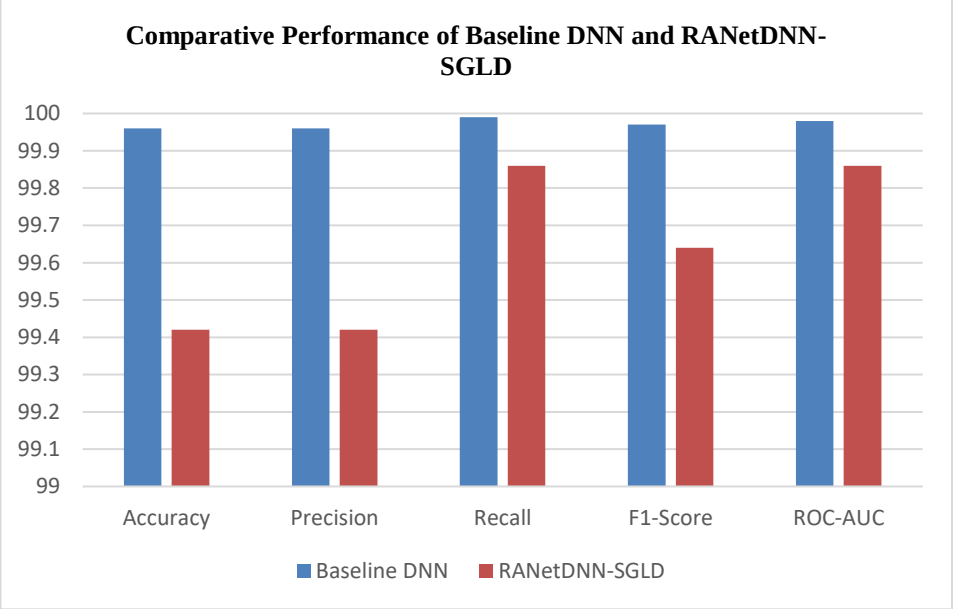

Figure 1. Comparative Performance of the Baseline DNN and Proposed RANetDNN-SGLD

Although both models achieved high classification performance on clean network traffic, the proposed RANetDNN-SGLD model provides an additional advantage through its adversarial training strategy and SGLD optimization. Conventional deep learning models often exhibit high accuracy under normal conditions but experience significant performance degradation when exposed to adversarially perturbed inputs. Therefore, evaluating the models under adversarial attack conditions is essential to determine their practical reliability in real-world SDN environments.

3.4 Adversarial Robustness Evaluation Using FGSM Attack

To evaluate the robustness of the proposed intrusion detection model against adversarial attacks, experiments were conducted using the Fast Gradient Sign Method (FGSM). FGSM generates adversarial examples by introducing carefully crafted perturbations to the input features, thereby challenging the model's ability to correctly classify network traffic. The perturbation magnitude was controlled using different epsilon (ε) values ranging from 0.001 to 0.050, where larger epsilon values represent stronger adversarial perturbations.

The robustness of the baseline Deep Neural Network (DNN) and the proposed RANetDNN-SGLD model was evaluated using the F1-score, as this metric provides a balanced assessment of precision and recall for intrusion detection tasks. The results of the robustness evaluation are presented in Table 4

Table 4. FGSM Robustness Evaluation of baseline DNN and RANetDNN-SGLD

Epsilon (ε)	Baseline DNN F1-Score	RANetDNN-SGLD F1-Score
0.001	0.9997	0.9946
0.005	0.9997	0.994
0.010	0.9926	0.9938
0.020	0.9921	0.9942
0.050	0.8294	0.9937

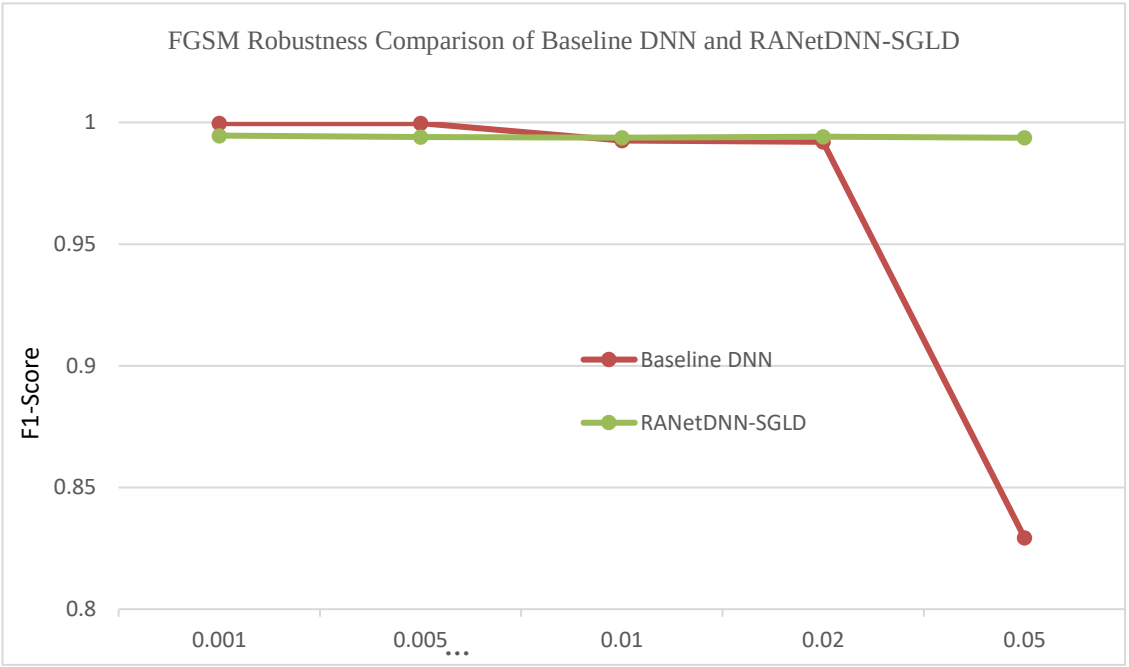

Figure 2. Comparison of the F1-Score of the Baseline DNN and RANetDNN-SGLD under Different FGSM Perturbation Levels

The results clearly demonstrate the superior adversarial robustness of the proposed RANetDNN-SGLD model. At lower perturbation levels (ε = 0.001–0.020), both the baseline DNN and the proposed model maintained high F1-scores, indicating effective intrusion detection under relatively weak adversarial attacks. However, as the perturbation strength increased, a major difference in performance became evident.

At ε = 0.050, the baseline DNN experienced a reduction in F1-score from 0.9997 to 0.8294, indicating that stronger adversarial perturbations degraded its detection capability. Whereas, the proposed RANetDNN-SGLD model maintained an F1-score of 0.9937, demonstrating only a negligible reduction in performance despite the increased attack intensity.

The superior robustness of the proposed model can be attributed to the integration of adversarial training and Stochastic Gradient Langevin Dynamics (SGLD) optimization. Adversarial training exposed the model to perturbed samples during learning, making it to identify malicious

patterns even under adversarial manipulation. Furthermore, the stochastic optimization process encouraged the model to converge toward flatter minima, improving generalization and reducing sensitivity to adversarial perturbations.

These findings demonstrate that the proposed RANetDNN-SGLD model not only achieves excellent intrusion detection performance under normal operating conditions but also maintains reliable performance when confronted with adversarial attacks. This is particularly important for real-world SDN deployments, where attackers may intentionally manipulate network traffic to evade intrusion detection systems.

4. Conclusion

This study presented an adversarially robust intrusion detection system for Software-Defined Networks by integrating a Resilient Adversarial Deep Neural Network (RANetDNN) with Stochastic Gradient Langevin Dynamics (SGLD) optimization, the experimental results demonstrated that the RANetDNN-SGLD model consistently maintained high intrusion detection performance while demonstrating higher robustness against FGSM adversarial attacks compared with the baseline Deep Neural Network.

Experimental evaluation using the InSDN dataset showed that the model achieved excellent classification performance under normal operating conditions and maintained strong robustness when evaluated against Fast Gradient Sign Method (FGSM) adversarial attacks. Comparative analysis showed that although both the baseline Deep Neural Network and the RANetDNN-SGLD model achieved high performance on clean data, the model showed greater resilience as the adversarial perturbation strength increased. In particular, and maintained high F1-score across all evaluated perturbation levels, whereas the performance of the baseline model deteriorated under stronger adversarial attacks. The integration of adversarial training with SGLD optimization contributed to improved model generalization and enhanced resistance to adversarial perturbations without compromising intrusion detection accuracy. These findings demonstrate that the proposed approach provides a reliable and effective solution for securing SDN environments against both predictable cyber threats and adversarial attacks.

Future research may explore the robustness of the framework against additional adversarial attack techniques, evaluate its performance on larger and more diverse SDN datasets, and explore its deployment in real-time Software-Defined Networking environments.

REFERENCES

Alyahyan, S. (2025). Enhancing Deep Learning with Resilient Adversarial Network (RANet): An Advanced Adversarial Resilience Training Framework for Robust Image Classification. *Neural Processing Letters*, *57*(4), 1–34. https://doi.org/10.1007/s11063-025-11777-3

Ataa, M. S., Sanad, E. E., & El-khoribi, R. A. (2024). Intrusion detection in software defined network using deep learning approaches. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-79001-1

Chaganti, R., Suliman, W., Ravi, V., & Dua, A. (2023). Deep Learning Approach for SDN-Enabled Intrusion Detection System in IoT Networks. *Information (Switzerland)*, *14*(1). https://doi.org/10.3390/info14010041

Doğan, S. M., ALKAN, M., KOÇAK, A., Koçak, A., & Alkan, M. (2025). Detection and mitigation of cyber-attacks in software defined networks using machine learning/deep learning: a systematic literature review, research challenges and future directions. *International Journal of Information Security*, *24*(5). https://doi.org/10.1007/s10207-025-01114-z

Elsayed, M. S., & Jurcut, A. D. (2021). *InSDN : A Novel SDN Intrusion Dataset*. https://doi.org/10.1109/ACCESS.2020.3022633

Hadi, M. R., & Mohammed, A. S. (2022). *A*

NOVEL A PPROACH TO N ETWORK I NTRUSION D ETECTION S YSTEM USING D EEP L EARNING FOR SDN : F UTURISTIC A PPROACH. 69–82. https://doi.org/10.5121/csit.2022.121106

Jawad, F. H. M. (2018). Software defined networking: An overview. *Journal of Engineering and Applied Sciences*, *13*(SpecialIssue13), 10790–10796. https://doi.org/10.36478/jeasci.2018.10790.10796

Li, L., Liu, J., & Wang, Y. (2024). *Geometric ergodicity of SGLD via reflection coupling*. *24*(5), 1–38. https://doi.org/10.1142/S0219493724500357

Mujahid, M., Mirdad, A. R., Alamri, F. S., Ara, A., & Khan, A. (2025). *Software de fi ned network intrusion system to detect malicious attacks in computer Internet of Things security using deep extractor supervised random forest technique*. https://doi.org/10.7717/peerj-cs.3103

Saini, S., & Chennamaneni, A. (2023). *A Review of the Duality of Adversarial Learning in Network Intrusion : Attacks and Countermeasures*. *May*.