Article

Detecting and Mitigating Adversarial Attacks: Model Comparison Between Tree-Based and Deep Learning-Based Intrusion Detection Systems

Idris Olanrewaju Ibraheem
https://orcid.org/0009-0002-3677-2478
Fountain University Osogbo, Nigeria
ibraheem.idris@fuo.edu.ng

Sanni-Akintunde Awaw Kehinde
https://orcid.org/0009-0006-4889-712X
Fountain University Osogbo, Nigeria
sanni-akintunde.awaw@fuo.edu.ng

Adewuyi M'Sahid Abiola
https://orcid.org/0009-0009-4764-8123
Al-Hikmah University Ilorin, Nigeria
msahid.adewuyi@outlook.com

Sanusi-Akintunde Aishat Taiwo
https://orcid.org/0009-0006-6594-6579
Federal University of Medicine and Medical Sciences, Abeokuta, Nigeria
akintundeaishattaiwo@gmail.com

Abstract

The study investigates the detection and mitigation of adversarial attacks targeting tree-based and deep learning-based intrusion detection systems (IDS). There has been a rise in advanced cyber threats; traditional IDS techniques often fall short, rendering them ineffective in some situations, prompting increased adoption of deep learning-based models as alternatives. These models are also prone to adversarial attacks that are creatively designed to deceive the system into classifying malicious traffic as benign. The Canadian Institute for Cybersecurity Intrusion Detection System 2017 (CICIDS2017) dataset was used for the study. Data were processed to handle missing values, normalisation features, and address the class imbalance using the ADASYN (adaptive synthetic) sampling approach. Evaluation was carried out on two models, namely the random forest classifier and the deep neural network system. An almost perfect accuracy of 99.80% was achieved with the random forest model, which shows its effectiveness against adversarial attacks like the fast gradient sign method (FGSM) and projected gradient descent (PGD). However, as promising as the deep neural network (DNN) system is, it only achieved an accuracy rate of 81%, which shows that it is open to adversarial attacks. Feature importance analysis highlighted critical network attributes, such as packet length and flow size, for intrusion detection. The finding shows the superior adversarial effectiveness of tree-based models relative to deep learning, which motivates for hybrid AI approaches and advanced future research. The study

UNISA university of south africa PRESS
Southern African Journal of Security
#20907 | 19 pages
https://doi.org/10.25159/3005-4222/20907
ISSN 3005-4222 (Online)
© Unisa Press 2026

emphasises the integration of adversarial training, interoperability of models, and detection of anomalies. The study contributes to the ever-evolving development of robust and secure IDS, while not overlooking the recurring issues in cybersecurity.

Keywords: adversarial attacks; deep learning security; robust neural networks; threat mitigation strategies; adversarial training

Introduction

The continuous evolution of cyber threats has prompted the development of advanced security mechanisms that protect digital infrastructures. For a long time, intrusion detection systems (IDS) have been playing a significant role in cybersecurity, with the provision of critical defence techniques against different activities of threat actors (Kuzior et al. 2024). Traditional IDS methods have historically relied on rule-based detection and some machine learning techniques, which often prove to be ineffective against advanced cyber-attacks (Mmaduekwe and Mmaduekwe 2024). Considering the ineffectiveness of traditional methods, deep learning-based IDS has emerged as a powerful alternative by incorporating advanced neural network architectures in the detection of complex attack patterns and anomalies in the traffic passing through the network (Smith, Lee, and Patel 2024).

Despite the effectiveness of deep learning-based IDS techniques, they are not impenetrable to threats. Adversarial attacks have always been known to pose persistent challenges to the systems because of how they carefully involve crafted patterns through which machine learning models are deceived into not classifying malicious traffic as an attack (Smith, Lee, and Patel 2024). These attacks manifest in several forms, including evasion attacks that bypass detections by modifying the input samples, and poisoning attacks that manipulate training data to degrade model performance over time (Smith, Lee, and Patel 2024). The vulnerabilities of deep learning models to these kinds of manipulation are of critical concern with regard to the dependability and strength of IDS in real-world cybersecurity applications.

This study investigated the issue of adversarial attacks and the effects they have on IDS that are built on deep learning techniques. The study provides a thorough review of using deep learning techniques for IDS, the impact of attacks, and the evolving protective apparatuses that are designed for model resilience enhancement. It also examined research gaps that are fundamental in the development of more secure and stronger deep learning-based IDS models in the future. By examining the known weaknesses of deep learning models and then implementing more efficient protective measures, cybersecurity professionals can strengthen the dependability of the IDS and protect the digital infrastructure from evolving threats. The study contributes to the evolving knowledge of cybersecurity through an analysis of these adversarial attacks' detection and mitigation techniques through a deep learning approach to achieve more trustworthy and resilient IDS deployments.

Literature Review

Deep Learning in IDS

The emergence of deep learning in the development of intrusion detection brings about a paradigm shift which enables advanced development of capable systems that understand and automatically extract patterns from a large amount of network traffic data. Traditional IDS, such as those that depend on rule-based methods or weak machine learning algorithms, have shortfalls when it comes to detecting new and emerging cyber threats. However, deep learning models like convolutional neural networks and recurrent neural networks have shown great capabilities in the identification of known and unknown attacks with more accuracy (Kimanzi et al. 2024).

According to Arjunan (2024), deep learning models are quite exceptional in handling the large amounts of data that are a fundamental part of network traffic datasets. Deep learning models can process raw traffic data without the need for extensive manual feature engineering, which substantially simplifies the processing pipeline. This adaptive capability renders deep learning particularly well-suited to the dynamic and heterogeneous nature of modern cybersecurity threats. For instance, Li et al.'s (2021) survey of deep learning applications in cybersecurity emphasised the efficiency of these models in the detection of intrusions. Their study shows how multidimensional, complex data can be taken into account using a deep learning architecture that is also able to detect new attack vectors, which makes deep learning applications important tools in stopping cybercrimes.

Mousa and Abdullah (2023) proposed a deep learning-based distributed denial of service (DDoS) detection system in which stacked autoencoders were used in the detection of anomalies in network traffic. It performed optimally, resulting in high detection rates and, at the same time, producing relatively low false positive rates. This is an example of how the enhancement of accuracy and stability in the operation of IDS can be done through the use of deep learning models. Basati and Faghih (2023) introduced a non-symmetric deep autoencoder (NDAE) for unsupervised feature learning in IDS. The NDAE model in their study uses zero-day attacks that used to be referred to as unknown subtle vulnerabilities in a service or protocol that attackers tried to exploit before either the intended system of attack finds a way to mitigate this attack vector, or the developer of the intended system of attack was able to patch the vulnerability. Their study shows the superiority of NDAE over classical machine learning approaches in such settings, providing more evidence of the advancement of deep architecture for stronger feature extraction (Mousa and Abdullah 2023).

Basati and Faghih (2023) also introduced a non-symmetric deep encoder for unsupervised feature learning in intrusion detection. The objective of the model is to address the challenges surrounding the detection of zero-day attacks, as the weaknesses it has are being exploited by threat actors before they are identified and mitigated. The study shows that NDAE performed better than traditional machine learning techniques,

which further shows the significance of the role played by deep learning in the enhancement of IDS capabilities.

Despite the progress that has been achieved, IDS built on the basis of deep learning do have their own challenges, the most pressing of which is proneness to adversarial attacks, which has reduced their effectiveness. Some of the deep learning models have weaknesses that are exploited by adversaries in order to bypass the established detection techniques, which poses a significant threat to the security of the network.

Adversarial Attacks on Deep Learning Models

The deployment of deep learning models has been met with adversarial attacks, which is a significant challenge when it comes to cyberthreat protection. Most of these attacks manipulate the data in a deliberate manner to deceive the deep learning models into making wrong predictions (Okafor 2024).

According to Goodfellow, Shlens, and Szeged (2014), who introduced the concept of adversarial examples in deep learning, a little manipulation of the input data can result in model misclassification with high confidence. As a result, Goodfellow, Shlens, and Szeged (2014) proposed the fast gradient sign method (FGSM) as a technique used for the generation of adversarial samples. Other researchers have investigated the application and implications of adversarial threats in different types of domains. Lourdu and Gunasekaran (2023) worked on loan eligibility prediction of adversarial samples, while Hassan et al. (2022) used the combination of fast gradient sign method and image processing methods for the creation of realistic adversarial samples for computer vision tasks. The study pointed out the fragile state of deep learning models, which resulted in more research interest in the development of stronger defence mechanisms in the mitigation of cyber threats.

Based on the foundation of deep learning and adversarial attacks, Goodfellow, Shlens, and Szegedy (2014) explored the portability of adversarial samples and found that an attack meant for a model can easily deceive other models as well. This finding reiterated the already established vulnerability of the deep learning system and stressed the need for a holistic defence strategy for protection against diverse and emerging threats and attack vectors. Carlini and Wagner (2016) proposed an advanced technique by taking adversarial attacks to another level entirely, such as the Carlini and Wagner (CandW) attack, which possesses the capability of bypassing existing protection measures. The study showed how dangerous adversarial attacks can be and the need to have in place more robust protective measures for deep learning-based IDS.

Adversarial attacks take many forms. In evasion attacks, for example, attackers modify the network traffic to avoid detection by the system. Through poisoning attacks, the attackers manipulate the training data used to train the IDS, and this leads to degraded model performance and reduced detection accuracy. Evasion and poisoning attacks are both threats to the integrity and effectiveness of deep learning-based IDS.

Defences Against Adversarial Attacks

To understand the increasing threat of adversarial attacks, researchers have developed several defence mechanisms to enhance the robustness and resilience of deep learning models. These can be classified into three main approaches: adversarial training, anomaly detection, and the model's ability to interpret (Kuzior et al. 2024; Papernot, McDaniel, and Goodfellow 2016).

Adversarial Training

One effective method to improve the scalability of deep learning models is through adversarial training to mitigate adversarial attacks by augmenting the training dataset with adversarial samples, which in turn expose the models to a wide range of potential scenarios of attacks while conducting the training. This approach gives the model an edge when handling adversarial inputs in a more realistic setting (Zhao 2024).

Madry et al. (2017) proposed a broader optimisation framework for adversarial training and showed how efficient it is against strong adversarial attacks. Their study also provided a theoretical understanding of the principles behind adversarial training and highlighted its potential to become a protective measure. Building on the work of Madry et al. (2017), Tramèr et al. (2017) worked on ensemble adversarial training, which banks on adversarial samples generated from many various models for further model enhancement. This creative but strong approach tried to address some of the shortfalls of traditional adversarial training, which shows a great advancement in the field of cybersecurity.

Anomaly Detection

Anomaly detection techniques are critical for identifying adversarial samples by detecting deviations from normal data patterns. These methods play a pivotal role in cybersecurity, where adversarial attacks often manifest as anomalous behaviour that can be flagged for further investigation and mitigation.

Recent advancements in anomaly detection have introduced novel approaches to enhance the robustness of deep learning models against adversarial threats. For instance, Mirzakhaninafchi (2024) developed a hybrid generative-adversarial framework that combines variational autoencoders (VAEs) with adversarial training to detect subtle anomalies in input data. Their approach leverages the reconstruction error of VAEs to identify inputs that deviate significantly from the learned distribution, providing a robust mechanism for distinguishing between legitimate and adversarial examples.

Another notable contribution was made by Zhang, Shi, and Huang (2024), who introduced a real-time anomaly detection system using a transformer-based architecture. Their model captures long-range dependencies in sequential data, making it particularly effective for detecting adversarial attacks in network traffic streams. The system achieves state-of-the-art performance in terms of detection accuracy and computational

efficiency, demonstrating its potential for deployment in large-scale cybersecurity applications.

Model Interpretability

The techniques that involve model interpretability aim to understand how deep learning models make decisions to identify which features are most vulnerable to adversarial attacks. Knowledge of how models make predictions can be used to develop more effective defence strategies.

In the last few years, considerable progress has been made in making deep learning models more interpretable. For example, Roshan and Zafar (2021) conducted a study that extended the SHAP (SHapley Additive exPlanations) framework to high-dimensional data, which is common in cybersecurity applications. The generalised SHAP approach combines domain-specific constraints to make feature contributions more explainable in order to get a more detailed view of model vulnerabilities.

Duan et al. (2023) proposed a local interpretable model-agnostic explanations-based approach for black-box adversarial attack (LIME-Attack), which uses LIME to identify discriminative regions for targeted perturbations, achieving high attack success rates with just a few deviations in models for adversarial defence. This method explains model predictions in the presence of adversarial perturbations. By simulating adversarial attacks during the explanation process, LIME-Attack gives more accurate and actionable insights into the weaknesses of deep learning models. In a recent study, Sharma et al. (2024) introduced a new framework called CausalSHAP, which is the combination of causal inference with SHAP values to understand the relationships between input features and model predictions. This also helps in identifying the pathways that are often used by adversarial attacks to enable the design of better defence mechanisms.

These studies show that interpretability is key to addressing adversarial attacks. By using advanced interpretability techniques, one can gain more insight into the inner workings of deep learning models and ultimately more secure and reliable systems against the evolving cyber threats.

Research Gaps and Challenges

Despite significant advances in adversarial defence for machine learning systems, several gaps persist in the practical deployment of these techniques in deep learning-based IDS. These challenges warrant critical examination beyond mere enumeration.

First, while adversarial training frameworks such as those proposed by Madry et al. (2017) and Tramèr et al. (2017) have demonstrated theoretical promise, their practical effectiveness is often limited by the specific attack types encountered during training. Models trained exclusively against FGSM or PGD (projected gradient descent)

perturbations may remain vulnerable to adaptive or black-box attacks, a limitation our results partially reflect. The 81% deep neural network (DNN) accuracy under adversarial conditions suggests that standard adversarial training alone is insufficient for real-world deployment.

Second, friction exists between interpretability and robustness. Tree-based models such as random forest offer inherently interpretable decision boundaries, which facilitate feature-level attribution (as our SHAP analysis illustrates); however, they lack the representational depth required to model complex temporal or sequential attack patterns. Deep architectures offer this representational power but at the cost of interpretability and, as this study demonstrates, adversarial robustness. This trade-off is not fully addressed in the existing literature and represents a fundamental challenge for hybrid model design.

Third, the generalisability of IDS research remains constrained by benchmark dataset dependency. The Canadian Institute for Cybersecurity Intrusion Detection System 2017 (CICIDS2017) dataset which was used for this study, while widely adopted, contains known class imbalance issues and reflects 2017 threat conditions. Recent adversarial attack vectors, such as adversarial patches in IoT environments or model inversion attacks, are not represented. This limits the direct applicability of comparative findings and calls for evaluation frameworks that span multiple, temporally diverse datasets.

These gaps motivate the present study's dual-model comparison under adversarial stress, and they inform the research directions proposed in the methodology section.

Methodology

Dataset and Preprocessing

Dataset

This study was carried out using the CICIDS2017 dataset, a network intrusion detection dataset containing labelled network traffic data consisting of normal and malicious activities. It includes different types of attacks, including DDoS, brute-force, infiltration, and botnet traffic, which makes it a good pick for the evaluation of IDS performance in challenges such as adversarial attacks.

Data Preprocessing

To have a strong dataset for training the model, the following preprocessing steps were used:

(1) Feature selection: the "Flow Duration," "Total Fwd Packets," and "Flow Packets/s" features were used due to their significance in the analysis of network traffic.

(2) Handling missing values: missing values were replaced using median imputation.

(3) Removing infinite values: infinite values were replaced with large finite values.

(4) Encoding categorical labels: the “Label” column was converted into numerical values using LabelEncoder.

Addressing Adversarial Attacks and Class Imbalance

(1) ADASYN (adaptive synthetic) sampling: for the handling of the class imbalance, ADASYN was incorporated to get synthetic samples of minority class so as to have a dataset that is balanced.

(2) Adversarial training: Attack strategies like FGSM and PGD were implemented to create more adversarial samples. The original and adversarial data were used to train the model for improved robustness.

Model Development and Training

The DNN was constructed as a fully connected feedforward architecture comprising four hidden layers with 256, 128, 64, and 32 neurons, respectively. Rectified linear unit activation functions were applied in all hidden layers to introduce non-linearity, while a Softmax activation was applied at the output layer to produce class probability distributions. The Adam optimiser was employed with a learning rate of 0.001 and a batch size of 64. To regularise the model and reduce overfitting, a dropout rate of 0.3 was applied after each hidden layer. The model was trained over 50 epochs with early stopping applied based on validation loss. All hyperparameters were initially set using domain heuristics and subsequently refined through GridSearchCV.

The random forest classifier was configured with 100 decision tree estimators (n_estimators=100), using Gini impurity as the splitting criterion and no maximum depth constraint, allowing trees to grow until leaf purity. GridSearchCV was used to tune the number of estimators and the minimum samples required at leaf nodes.

Findings

The results indicate that the random forest classifier significantly outperformed the DNN in both accuracy and adversarial robustness. The random forest model demonstrated strong generalisation capabilities, achieving an F1-score of 1.00, while also exhibiting resilience against FGSM and PGD attacks. However, the DNN, despite its potential for deeper pattern recognition, displayed some weakness to adversarial concerns with an overall lower accuracy of 81%.

The analysis of the feature importance shows the critical network attributes that were used for the classification, and in turn offers valuable insight into the future optimisation of intrusion detection. Future research can explore and hybridise AI methods that

combine the deep feature extraction of neural networks' power and the interoperability of tree-based models for effective enhancement of cyberthreat mitigation.

Discussion

The study examined and analysed experimental results that were obtained while implementing the random forest classifier and DNN for the detection of intrusions. The evaluation carried out included performance metrics for models, resilience of adversarial attacks, and graphical illustration for classification.

Model Performance

Standard classification metrics such as F1-score and area under the receiver operating characteristic curve (AUC-ROC) were used to assess the effectiveness of the models, and the results show that for network anomalies detection, the random forest classifier is better, while also pointing out the challenges that DNNs face.

Random Forest Classifier

During the training and testing phases, the random forest classifier exhibited outstanding performance, and 99.99% accuracy was achieved on the dataset training. The model achieved an almost perfect classification. While an impressive 99.80% accuracy was achieved on the test dataset, this feat shows excellent generation capabilities. The model achieved a balanced performance, as it excelled in precision and recall with an F1 score of 1.00.

The near-perfect accuracy of 99.80% and F1-score of 1.00 achieved by the random forest classifier merit careful adaptation. While fivefold cross-validation was applied to mitigate overfitting, the unique nature of these metrics warrants acknowledgement of two potential confounds. First, the CICIDS2017 dataset contains known temporal locality in its traffic captures; if training and evaluation folds are not stratified temporally, information from later captures may implicitly inform earlier predictions, constituting a form of data leakage. Second, the limited feature set employed (three primary features) may reduce the model's exposure to noisy or adversarially perturbed attributes, artificially inflating apparent robustness. These considerations do not invalidate the findings, but they underscore the importance of replication on a fully held-out, temporally stratified test partition in future work.

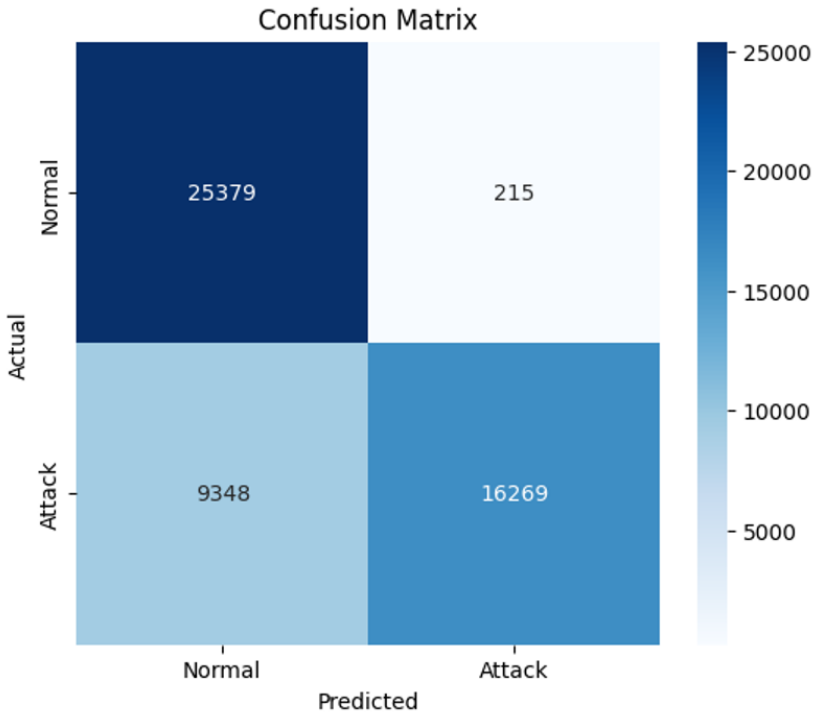

Figure 1: Confusion matrix (normal vs. attack classification)

Figure 1 presents the confusion matrix for the random forest classifier, displaying the distribution of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) across the normal and attack traffic classes. The matrix reveals a highly skewed distribution towards correct classifications, with near-zero FP and FN rates. From an operational standpoint, the near-zero FN rate is particularly significant: it implies that the model rarely allows malicious traffic to pass undetected, the most notable failure mode in a network intrusion detection context. The low FP rate is equally important for deployment, as excessive false positives impose a substantial analyst workload and can lead to alert fatigue. These results suggest that the random forest classifier achieves a practically acceptable operating point for network security deployment. However, the reliance on three features for classification introduces a risk of failure under distributional shift; future refinements incorporating a broader feature set should be validated against these confusion matrix baselines to ensure that expanded feature representations do not introduce additional FP inflation.

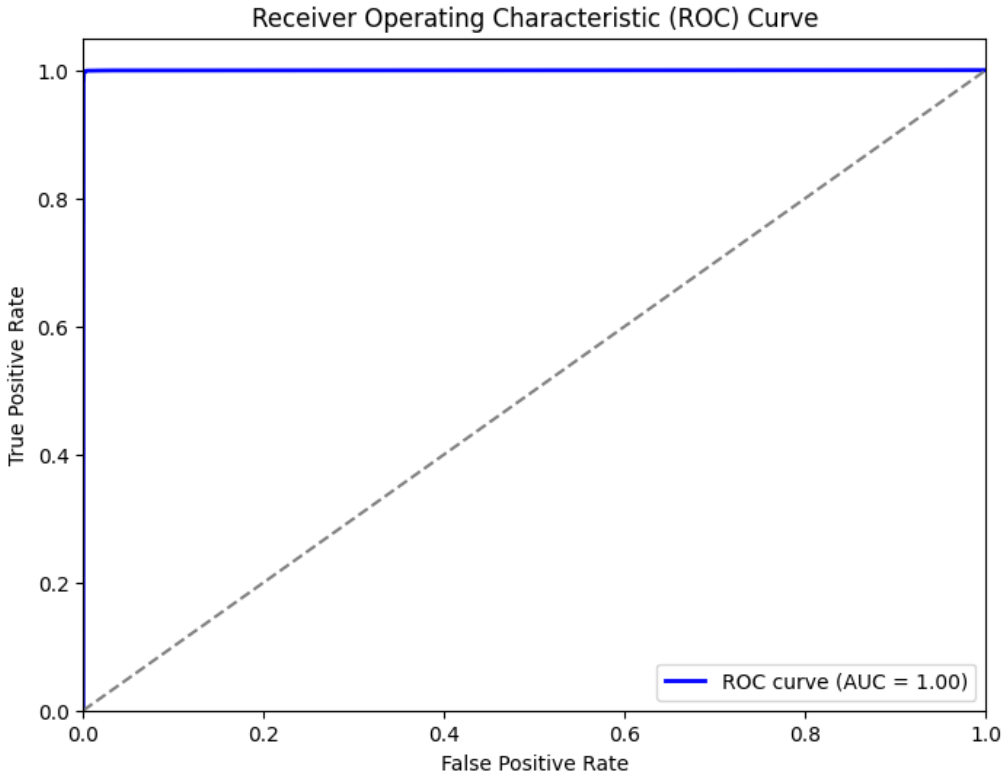

Figure 2: ROC curve based on model performance evaluation

Figure 2 presents the receiver operating characteristic (ROC) curve for the random forest classifier. The ROC curve plots the TP rate (sensitivity) against the FP rate (1-specificity) across the full range of classification thresholds, providing a threshold-independent measure of discriminatory power. The area under the curve (AUC) of 0.99 indicates near-perfect class separability, confirming that the model maintains strong intrusion detection performance across virtually all operating thresholds. This is of direct relevance to cybersecurity practitioners who often operate models at custom thresholds calibrated to their specific tolerance for FP versus FN. The near-unity AUC further validates the confusion matrix results, corroborating that the classifier's high accuracy is not an artefact of a single favourable threshold. Nonetheless, the extremely high AUC should be interpreted alongside the caveat on data leakage discussed under model performance. The AUC values approaching 1.00 on benchmark datasets are not always reproducible in live deployment environments where class distributions and attack profiles differ from training conditions.

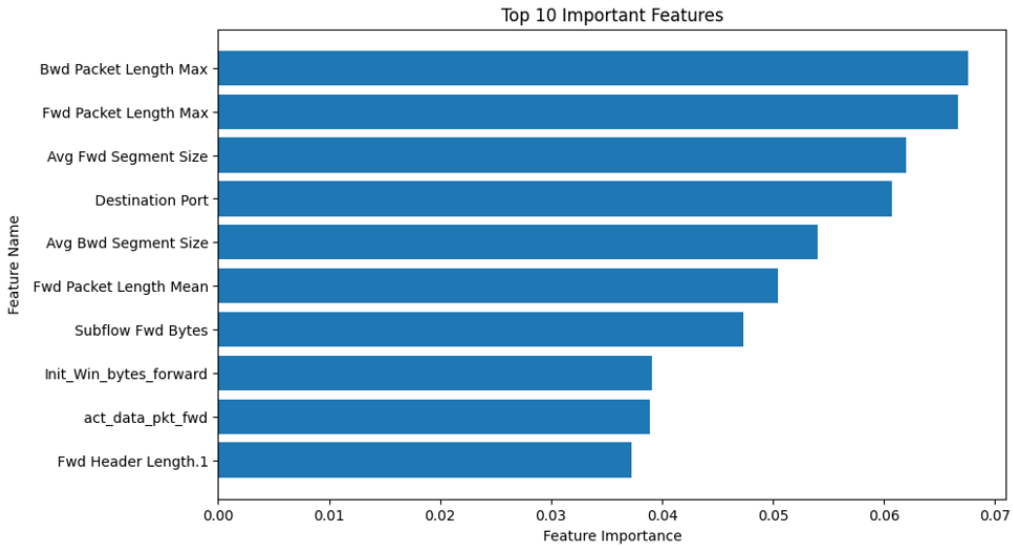

Figure 3: Feature importance

Figure 3 presents a bar chart of the top 10 most discriminative features identified by the random forest classifier, ranked by their Gini-based importance scores. Backward packet length maximum (Bwd Packet Length Max) and forward packet length maximum (Fwd Packet Length Max) emerge as the two most important features. This finding has a clear network security interpretation: attack traffic, particularly DDoS and flooding attacks, frequently generates anomalous packet size distributions relative to normal browsing or enterprise traffic. Large maximum packet lengths in the backward direction (server-to-client) can indicate scanning or data exfiltration, while forward direction anomalies often reflect payload injection or command-and-control traffic. Flow-based metrics and destination port numbers also figure prominently, which is consistent with the literature on network flow analysis for intrusion detection (Li et al. 2021). Crucially, these features are precisely the attributes most susceptible to adversarial perturbation: an adversary who understands the model's reliance on packet length statistics can craft traffic that mimics benign packet size distributions while embedding malicious payloads. This observation reinforces the argument for adversarial training and input transformation defences and provides concrete guidance for feature selection in next-generation IDS development.

Deep Neural Network

The DNN was trained over 50 epochs, achieving an accuracy of 81% on the test dataset. The model showed promising results, although its performance was considerably lower than that of the random forest classifier. DNN requires further optimisation, such as hyperparameter tuning, additional feature engineering, or more extensive training on augmented datasets, due to a few discrepancies in the model.

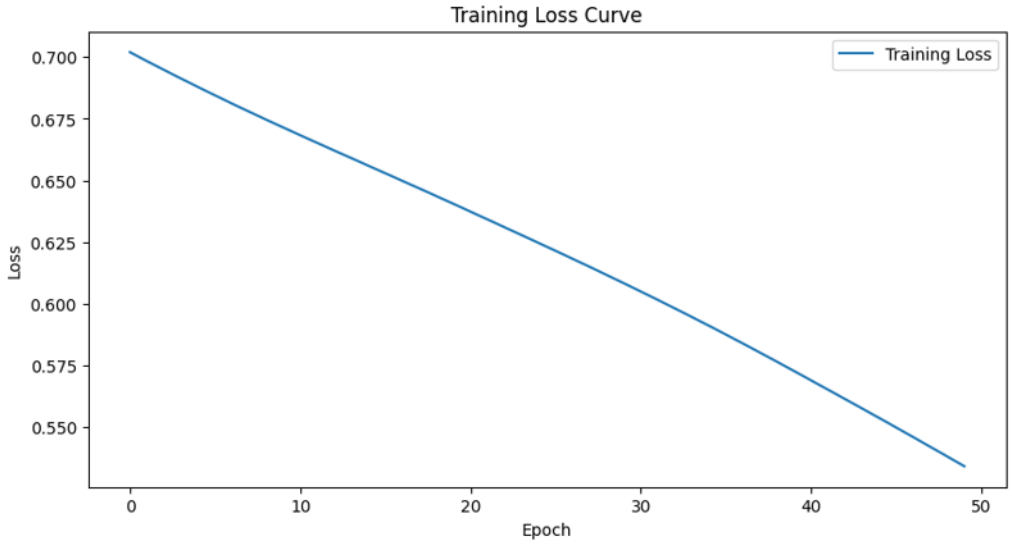

Figure 4: Training loss curve for function decrement

The training loss curve in Figure 4 shows how the loss function decreases over epochs. The steady decline in loss indicates effective learning, as the model improves its predictions with more training iterations. If the loss continues to decrease without signs of overfitting, the model is learning the underlying patterns in the data effectively. The training loss curve illustrates the learning progression of the DNN. The model was effective in the learning patterns within the data as shown by the continuous decrease in loss across epochs. However, the model would benefit from further improvements like dropout regularisation or batch normalisation, which may be good for the enhancement of the model's ability to prevent overfitting.

Adversarial Attack Mitigation

Given the increase in much stronger cyber threats, the resilience of both models was tested against adversarial perturbations, specifically using FGSM and PGD attacks.

Impact of FGSM and PGD Attacks

Random forest classifier: The model exhibited notable resilience against adversarial attacks. Since decision trees rely on discrete splitting mechanisms, small variations in input data had a minimal impact on classification results. This significant efficiency makes random forest a better choice in mitigating adversarial-resilient IDS.

DNN: Compared with the performance of random forest, the DNN suffered a significant decrease in performance in the detection of adversarial attacks. Continuous-valued transformations' reliance on the model made it prone to small but carefully crafted variations, leading to misclassifications. This shows the vulnerability of deep learning models in cybersecurity applications.

More protective techniques like adversarial training and defensive distillation can improve the strength and scalability of DNN against adversarial threats. Future studies should focus on integrating such techniques to enhance the robustness of AI-driven network security models.

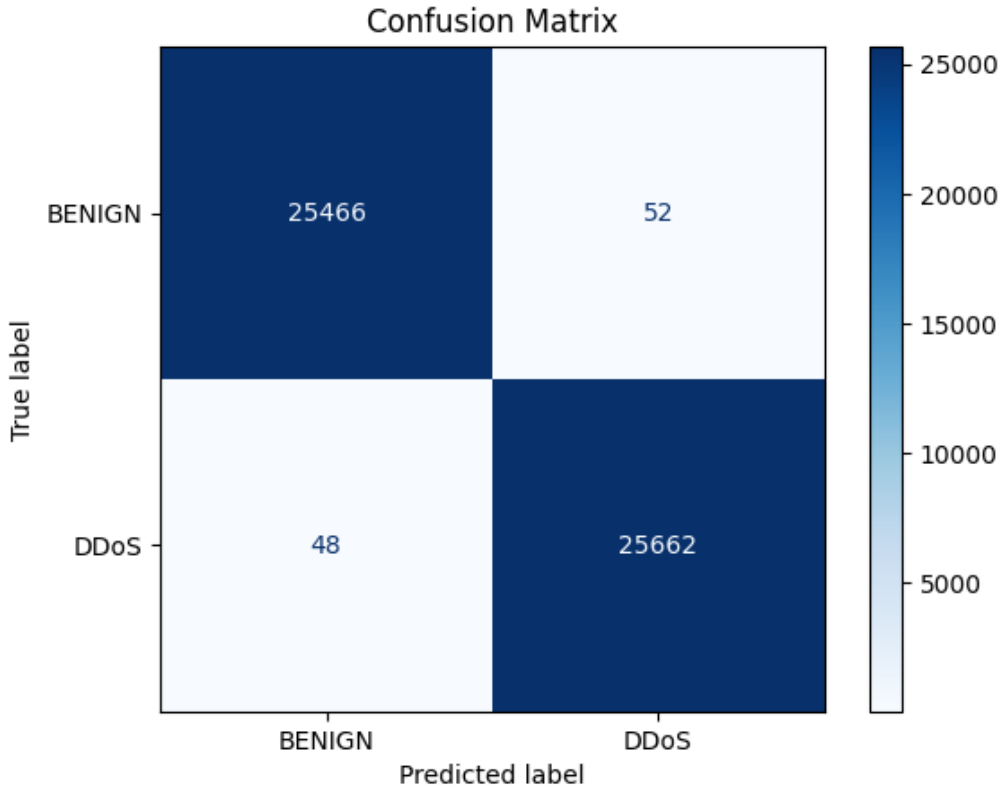

Figure 5: Confusion matrix for (benign vs. DDoS classification)

Figure 5 shows a confusion matrix for distinguishing between benign and DDoS traffic. The model demonstrates high accuracy, as there are minimal misclassifications. The reduced number of false positives and false negatives shows that the performance of the classifier is optimal in the detection of DDoS attacks with few errors.

SHAP Interaction Analysis of Key Network Traffic Features

Figure 6 presents the SHAP interaction plot, which provides insights into how key network traffic features, specifically destination port and flow duration, contribute to the model's decision-making process in detecting adversarial attacks. The distribution of SHAP interaction values through the x-axis shows the extent to which these features influence the predictions when assessed together. The presence of both red and blue points indicates varying impacts across different classes, which corresponds with normal and attack traffic. The model appears to assign different interaction strengths

based on the specific feature values, reinforcing the importance of these attributes in distinguishing between benign and malicious network traffic. This visualisation aligns with our study's objective of understanding adversarial efficiency, as it indicates which features play a crucial role in mitigating adversarial perturbations and improving classification reliability.

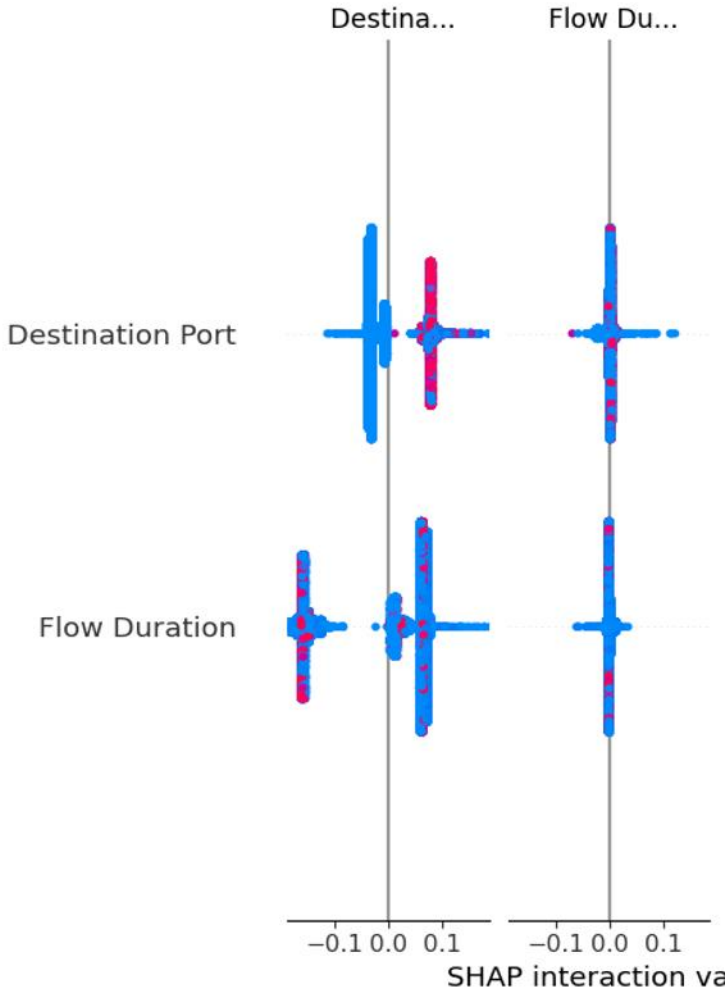

Figure 6: SHAP interaction plot for destination port and flow duration

Future Work and Research Directions

While this study enhances intrusion detection accuracy and robustness, several areas warrant further research. The weakness of deep learning models to adversarial attacks shows the need for more advanced defensive distillation, adversarial training, and feature transformation techniques. Additionally, future studies should focus on the combination of machine learning and deep learning approaches to develop a hybrid

model that can optimise accuracy and efficiency by employing ensemble methods and federated learning for enhanced security. Future research should also focus on real-time deployment and scalability, which ensures IDS models are able to handle large-scale network traffic efficiently. Expansion of the models by testing them on diverse datasets and optimising computational efficiency will improve real-world applicability. The model can be enhanced further by including explainable AI (XAI), such as SHAP and LIME, for model interpretability, which enables security analysts to understand AI-driven decisions.

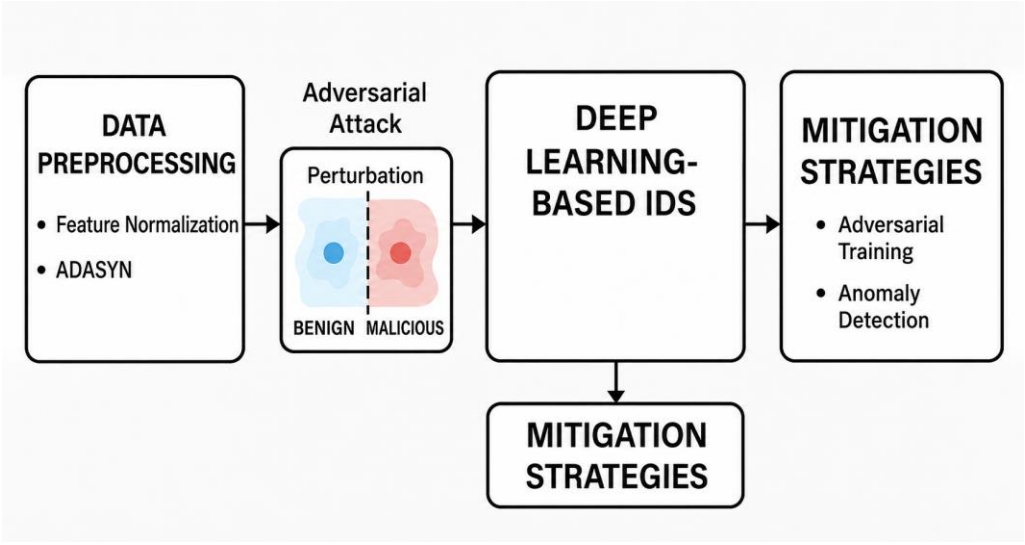

Figure 7: Overview of the proposed AegisDL framework

The proposed AegisDL framework, as seen in Figure 7, is a conceptual architecture for future research. It integrates data preprocessing, adversarial attack modelling, deep learning-based intrusion detection, and mitigation strategies to improve the robustness and resilience of intrusion detection systems against adversarial attacks. Future work will involve implementing, validating, and evaluating the framework using benchmark intrusion detection datasets and adversarial attack scenarios.

Conclusion

This study highlights the vulnerabilities of deep learning-based IDS to adversarial attacks and proposes effective mitigation strategies. Using the CICIDS2017 dataset, we demonstrated that the random forest classifier performs better than the DNN in both accuracy (99.80%) and adversarial robustness, demonstrating its resilience against FGSM and PGD attacks. However, even when the DNN possesses the capability of deep pattern recognition, it also shows signs of significant proneness to adversarial misclassification as evidenced by the 81% accuracy score. Packet length and flow size metrics are important in differentiating between benign and attack-intended traffic. The

findings reiterate the need for advanced defence mechanisms like anomaly detection and adversarial training to enhance the efficiency of IDS. Future research can enhance the model further by working on hybrid AI models that use a combination of the interpretability of tree-based deep learning models to address evolving cyber threats and also enable security analysts to understand the patterns of adversarial attacks.

References

Arjunan, T. 2024. “Real-Time Detection of Network Traffic Anomalies in Big Data Environments Using Deep Learning Models.” *International Journal for Research in Applied Science and Engineering Technology* 12 (III): 844–850. https://doi.org/10.22214/ijraset.2024.58946

Basati, A., and M. M. Faghih. 2023. “APAE: An IoT Intrusion Detection System Using Asymmetric Parallel Auto-Encoder.” *Neural Computing and Applications* 35 (7): 4813–4833. https://doi.org/10.1007/s00521-021-06011-9

Carlini, N., and D. A. Wagner. 2016. “Towards Evaluating the Robustness of Neural Networks.” In *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. Piscataway, NJ: IEEE. https://doi.org/10.1109/SP.2017.49

Okafor, M. O. 2024. “Deep Learning in Cybersecurity: Enhancing Threat Detection and Response.” *World Journal of Advanced Research and Reviews* 24 (3): 1116–1132. https://doi.org/10.30574/wjarr.2024.24.3.3819

Duan, Y., X. Zuo, H. Huang, B. Wu, and X. Zhao. 2023. “A Local Interpretability Model-Based Approach for Black-Box Adversarial Attack.” In *Data Mining and Big Data, DMBD 2023*, edited by Y. Tan and Y. Shi, 3–15. Singapore: Springer Nature. https://doi.org/10.1007/978-981-97-0844-4_1

Goodfellow, I. J., J. Shlens, and C. Szegedy. 2014. “Explaining and Harnessing Adversarial Examples.” Preprint, submitted December 14, 2014. https://doi.org/10.48550/arXiv.1412.6572

Hassan, M., S. Younis, A. Rasheed, and M. Bilal. 2022. “Integrating Single-Shot Fast Gradient Sign Method (FGSM) with Classical Image Processing Techniques for Generating Adversarial Attacks on Deep Learning Classifiers.” In *Proceedings of the SPIE 12084, Fourteenth International Conference on Machine Vision (ICMV 2021), 1208415 (5 March 2022)*. https://doi.org/10.1117/12.2623585

Kimanzi, R., P. Kimanga, D. Cherori, and P. K. Gikunda. 2024. “Deep Learning Algorithms Used in Intrusion Detection Systems – A Review.” Preprint, submitted February 23, 2024. https://doi.org/10.48550/arXiv.2402.17020

Kuzior, A., I. Tiutiunyk, A. Zielińska, and R. Kelemen. 2024. “Cybersecurity and Cybercrime: Current Trends and Threats.” *Journal of International Studies* 17 (2): 220–239. https://doi.org/10.14254/2071-8330.2024/17-2/12

Li, G., P. Sharma, L. Pan, S. Rajasegarar, C. Karmakar, and N. Patterson. 2021. “Deep Learning Algorithms for Cyber Security Applications: A Survey.” *Journal of Computer Security* 29 (5): 447–471. https://doi.org/10.3233/JCS-200095

Lourdu Mahima Doss P., and M. Gunasekaran. “Generating and Defending Against Adversarial Examples for Loan Eligibility Prediction.” In *2023 International Conference on System, Computation, Automation and Networking (ICSCAN)*, 1–6. Piscataway, NJ: IEEE. https://doi.org/10.1109/ICSCAN58655.2023.10395585

Madry, A., A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. 2017. “Towards Deep Learning Models Resistant to Adversarial Attacks.” Preprint, submitted June 19, 2017. https://doi.org/10.48550/arXiv.1706.06083

Mirzakhaninafchi, H. 2024. “Comparative Analysis of Deep Learning-Based Anomaly Detection Models for GPS Spoofing Detection.” MA thesis, South Dakota State University.

Mmaduekwe, U., and E. Mmaduekwe. 2024. “Cybersecurity and Cryptography: The New Era of Quantum Computing.” *Current Journal of Applied Science and Technology* 43 (5): 41–51. https://doi.org/10.9734/cjast/2024/v43i54377

Mousa, A. K., and M. N. Abdullah. 2023. “An Improved Deep Learning Model for DDoS Detection Based On Hybrid Stacked Autoencoder and Checkpoint Network.” *Future Internet* 15 (8): 278. https://doi.org/10.3390/fi15080278

Papernot, N., P. McDaniel, and I. Goodfellow. 2016. “Transferability in Machine Learning: From Phenomena to Black-Box Attacks Using Adversarial Samples.” Preprint, submitted May 24, 2016. https://doi.org/10.48550/arXiv.1605.07277

Roshan, K., and A. Zafar. 2021. “Utilizing XAI Technique to Improve Autoencoder Based Model for Computer Network Anomaly Detection with Shapley Additive Explanation (SHAP).” Preprint, submitted December 14, 2021. https://doi.org/10.48550/arXiv.2112.08442

Sharma, N. A., R. R. Chand, Z. Buksh, A. B. M. Ali, A. Hanif, and A. Beheshti. 2024. “Explainable AI Frameworks: Navigating the Present Challenges and Unveiling Innovative Applications.” *Algorithms* 17 (6): 227. https://doi.org/10.3390/a17060227

Smith, J., A. Lee, and R. Patel. 2024. “Adversarial Attacks on Deep Learning-Based Intrusion Detection Systems: A Review.” *Computers and Security* 115: 102195. https://doi.org/10.1016/j.cose.2024.102195

Tramèr, F., A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. 2017. “Ensemble Adversarial Training: Attacks and Defenses.” Preprint, submitted May 19, 2017. https://doi.org/10.48550/arXiv.1705.07204

Zhang, M., X. Shi, and J. Huang. 2024. “Real-Time Anomaly Detection in Time Series Using Transformer-Like Architecture.” In *2024 IEEE International Conference on Artificial Intelligence Testing (AITest)*, 150–151. Piscataway, NJ: IEEE. https://doi.org/10.1109/AITest62860.2024.00026

Zhao, Y. 2024. “Improvement of the Robustness of Deep Learning Models Against Adversarial Attacks.” *Applied and Computational Engineering* 75: 285–289. https://doi.org/10.54254/2755-2721/75/20240556