Review of Network Intrusion Systems Evaluated on NSL-KDD and CIC-IDS2017 Datasets

Intan Tahira Binti Zamri
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
tp067076@apu.edu.my

Julia Juremi
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
julia.juremi@apu.edu.my

***Abstract*— The increase in the role of information technology in day-to-day tasks inevitably calls for heightened security measures, so as to protect sensitive data from falling into the wrong hands. The evolution of computer systems has incentivised the spawning of even more advanced and dangerous types of cyber-attacks, making it more of a challenge for security systems to identify them in an efficient and accurate manner. Network Intrusion Detection Systems (NIDSs), mostly operating on anomaly-based detection schemes, must comprise machine learning frameworks which are robust enough to effectively detect most network attack groups. Recent studies which focus on building efficient NIDSs present an amalgamation of techniques, from the stage of data pre-processing, feature selection to classification, each having its own strengths and limitations. This paper reviews several existing NIDS models which have been evaluated on benchmark datasets NSL-KDD and CIC-IDS2017. For both datasets, genetic algorithm-implemented models outperformed all other models across most performance metrics.**

Keywords*— *Cyber-attacks, Machine Learning, Network Intrusion Detection Systems, Anomaly-Based Detection, Feature Selection, Classification, NSL-KDD, CIC-IDS2017

I. INTRODUCTION

Due to the increase in computer and information technology usage in our daily lives, there is higher demand for means so as to secure and protect our data using sophisticated network security mechanisms. A Network Intrusion Detection System (NIDS) is a vital unit of any information system, as its purpose is to identify malicious activity in a network through monitoring and analysis of the network traffic behaviour (Scarfone and Mell, 2010). In the absence of a sturdy NIDS, a cyber-attacker may acquire sensitive information from a network system, causing a disruption to its operations. According to a security review presented in the 2020 National Technology Security Coalition, cyber-threats emerged more vigorous every month for that year (National Technology Security Coalition, 2020). This brings more concern with the emergence of newer, more sophisticated attacks, due to rapid evolvement of network techniques, hence, the rise in research for building effective NIDS algorithms.

Anomaly detection, a classification task in supervised machine learning (ML), is a form of data analysis which has been applied in in many existing NIDSs. Also known as outlier detection or deviation detection, this task involves identifying data points in a dataset which are anomalous from other data points, utilising feature vectors as a basis. Data anomalies are valuable towards research in a myriad of application domains, such as studying unusual traffic patterns in a network, irregular behaviour in credit card transactions, anomalies in medical imaging, and many more (Ahmed et al., 2016). Network anomalies deviate from the standard system operation, compromising system efficiency, and to some extent, result in system paralysis.

There are several datasets publicly available for researchers to evaluate their proposed NIDS algorithms. The NSL-KDD and CIC-IDS2017 datasets are fair benchmarks for researchers due to their close mimic of real-word internet traffic. Therefore, the proposed models reviewed in this paper will be of those which used either one of these datasets in their analysis.

In Section II, some preliminaries related to this topic are discussed, namely some introduction to the two benchmark cybersecurity datasets of interest (NSL-KDD and CIC-IDS2017), network attack types and model performance metrics. In Section III, a review of several existing models proposed to construct a NIDS is presented, which are categorised based on the dataset used for evaluation. Section IV concludes this paper, along with suggestions for future research.

II. PRELIMINARIES

A. NSL-KDD

The NSL-KDD dataset is useful for comparing NIDS models as its training and test sets are ratioed considerately such that researchers need not sample the dataset separately, and consistent comparisons can be made as well. The records available in this dataset are also sufficient in order to train a model into a robust NIDS. This dataset offers two segments; one meant for model training, with 25,192 and 125,973 data instances in each, and another two for model evaluation, with 22,543 and 11,850 data instances in each. (Canadian Institute for Cybersecurity, 1999). Table I shows the available NSL-KDD datasets.

TABLE I. AVAILABLE NSL-KDD DATASETS AND THEIR NUMBER OF INSTANCES (ADAPTED FROM SING AND KHARE (2021))

NSL-KDD DATASET	NO. OF DATASET INSTANCES
NSLKDDTRAIN20P	25,192
NSLKDDTRAIN+	125,973
NSLKDDTEST+	22,543
NSLKDDTEST21	11,850

This dataset has four categories of attacks; Denial of Services (DoS), Probe, Remote to Local (R2L), User to Root

(UR), and the fifth label being Normal. The descriptions of these attacks can be found in Section II(c). More than half of the records are labelled Normal, with DoS being the second most common class in the dataset, and the rest having the lowest percentages. This is accurately in line with the distribution of cyber-attacks proliferating today. (Canadian Institute for Cybersecurity, 1999)

B. CIC-IDS2017

The generation of the CIC-IDS2017 dataset is inspired by lifelike background network traffic using the most relevant and frequent attack types which resemble real-word network data. It comprises labeled network flows and data which was captured periodically over a span of 5 days. The attacks cover an additional variety of categories such as Heartbleed, Infiltration, Botnet and several more. To mimic human interactions and naturalistic benign background traffic, a B-Profile system (Sharafaldin et al., 2017) was utilised. (Canadian Institute for Cybersecurity, 2017)

C. Network Attack Types

Four common categories of network attacks are as shown in Table II. These categories are seen in relatively smaller datasets such as KDD99, which contains 40 network behaviours categorised into five categories, the fifth one being Normal.

TABLE II. NETWORK ATTACK CATEGORIES AND THEIR DESCRIPTIONS (ADAPTED FROM PAJOUH ET AL. (2017))

Network Attack Category	Description
Denial of Services (DoS)	Authorised users are denied service usage
Probe	Attacker attempts to obtain confidential data about the target host
Remote to Local (R2L)	Attacker attempts to remotely break into the victim machine
User to Root (U2R)	Using already-attained local access to target machine, attacker attempts to gain rights that of a super user

There are many other network attack types such as Infiltration, Web Attack and Bot, which can be seen in larger datasets such as CIC-IDS2017. An efficient NIDS should be able to correctly classify most of the network attack types within an acceptable amount of computational time.

D. Model Performance Metrics

The performance of NIDS models can be measured using the metrics such as Accuracy, Precision, Recall, F1-Score and Area under Curve, which are elaborated in Table III.

TABLE III. MODEL PERFORMANCE METRICS AND THEIR EQUATIONS (ADAPTED FROM HOSSIN AND M.N. (2015))

Performance Metric	Equation
Accuracy (Acc)	(TP + TN)/ (TP + TN + FP + FN)
Precision (Prec)	TP / (TP + FP)
Recall (Re)	TP / (TP + FN)
F1-Score (F1)	2 / (1/Precision + 1/Recall)
Area Under Curve (AuC)	Area under ROC Curve (Plot of TPR against FPR)

TP, TN, FP, FN, TPR, FNR are described in Table IV .

TABLE IV. MODEL PERFORMANCE METRICS AND THEIR COMMENTS (ADAPTED FROM HOSSIN AND M.N. (2015))

Performance Metric	Description or Equation
True Positive (TP)	Normal network behavior correctly classified as normal
True Negative (TN)	Anomalous network behavior correctly classified as anomalous
False Positive (FP)	Anomalous network behavior falsely classified as normal
False Negative (FN)	Normal network behavior falsely classified as anomalous
True Positive Rate (TPR)	TP / (TP + FN)
False Negative Rate (FNR)	FN / (TP + FN)
True Negative Rate (TNR)	TN / (TN + FP)
False Positive Rate (FPR)	FP / (TN + FP)

Researchers carry out model evaluation of their proposed NIDS algorithm using if not all, some of the performance metrics mentioned above. The 'Recall' metric indicates the detection rate, which its inclusion in model evaluation should be a standard procedure in order to measure the true performance of a NIDS model as well as to ease comparison between different frameworks.

III. RECENT WORK

Numerous studies have been carried out as efforts to develop efficient NIDSs. Some recently proposed models built for NIDSs which were evaluated using benchmark datasets CIC-IDS2017 and NSL-KDD will be looked into in this section, with one additional review of a model evaluated on the IDS2017 dataset.

A. CIC-IDS2017 Dataset

Ding and Li (2022) proposed a model for a NIDS which exploits a graph convolution network (GCN), a long short-term memory (LSTM) network and a soft attention mechanism. The GCN is purposed to learn hidden relationships within a static network during standard operations, such as the interaction between network nodes with outbound traffic and between the network nodes themselves. The LSTM network traces the change in network traffic over time, and the attention mechanism is integrated in the output of the LSTM to focus on the most relevant features. At the end, a SoftMax function is utilised to obtain the classification results. In addition to the CIC-IDS2017 dataset, their also evaluated their model using the CTU-13 dataset. During evaluation, their model had improved detection rates for all attack types, as it can detect low-intensity attacks which do not vary much statistical values. Furthermore, since the model has better feature learning capabilities from utilization of unseen network traffic relationships, it showed a superior detection rate on CIC-IDS2017, which is a larger and more complex dataset than CTU-13.

Chen et al. (2022) proposed a NIDS model which operates on deep belief networks (DBNs) and LSTM networks. They pointed out that machine learning models highly rely on dataset features, sample distribution and the sample amount, which limits the extent of their performance to datasets which are uniformly distributed, have adequate samples and well-selected features. Therefore, to address real-time cyber threats, a NIDS would require orderly updates. This led to their focus on effective feature extraction and enhancement of scalability towards the emergence of unseen attacks. The feature extraction phase aims to shrink the dimension of the raw data using a DBN made of several Restricted Boltzmann

Machines (RBMs), followed by a LSTM network to learn the time sequence information in the data, and perform classification of the data. The KDD99 and CIC-IDS2017 datasets were used in their model evaluation. Their results showed improvement in accuracies for both datasets after implementing the DBN. However, their proposed method showed inadequacy in detecting the minority of categories.

Aksu and Aydin (2022) proposed a NIDS which leverages a modified genetic algorithm (MGA) for feature selection. Their study also included intrusion detection for in-vehicle networks. The problems focused on in their research were long computation times and low detection performance due to redundant features, variation of performance by the classifier based on varying combinations of classifiers and features, and emergence of unknown attacks. Their modified genetic algorithm was based on k-fold cross validation (CV), with a wrapper approach used in feature selection. To avoid overfitting, a simple hold-out approach and 10-fold cross-validation was used. The MGA m-feature selection stage analysed the consistency of the reduced feature subsets. After obtaining a candidate feature subset, five classifiers; Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbours (K-NN) and Linear Discriminant Analysis (LDA) were trained and tested on the dataset, with the best classifier chosen to build their NIDS. As part of evaluation, they compared the models' performances before and after applying feature selection on datasets HCRL-car hacking, UNSWNB15 and CICIDS2017, and the feature selection was found to improve all performance metrics and reduce computational time for all classifiers, except for LDA and LR. The DT classifier showed the highest performance metrics, making it the final used classifier for their NIDS. Using only 5, 7 and 9 features, their proposed model could detect intrusions with 100% accuracy compared to peer methods which used many more features, yet, showed poorer performance.

Panigrahi et al. (2022) proposed a hybrid NIDS model using Naïve Bayes and Decision (DTNB) and a Multi-Objective Evolutionary Feature Selection (MOEFS) technique. As part of pre-processing, reciprocal down-sampling was done on the CIC-IDS2017 dataset. Their feature selection stage extracted 5 features from the dataset, and these features were sent to the DTNB hybrid classifier. The performance metrics used to measure the model's performance were Acc, FPR, Prec and AuC. When compared to peer techniques, the proposed model could detect 14 attack classes using only 5 features, compared to others which focused on around 5 attacks. The given imbalanced dataset was also handled well through their down-sampling technique. The DTNB's consistent accuracy, precision and detection rate proved its versatility by using only 5 features, giving enough evidence that the DTNB hybrid is robust in detecting cyber-attacks.

B. NSL-KDD Dataset

Pajouh et al. (2017) used Naïve Bayes (NB) as their ML technique for their NIDS model. They integrated LDA to extract the most relevant features from NSL-KDD dataset, followed by a bucketing approach for storing the reduced dataset using a K-Dimensional tree data structure. K-NN classifiers were used as their certainty factor voting version. The performance of their model indicated that through a certainty factor (CF) for similarity measure, an acceptable detection rate against rare and dangerous attacks such as U2R and R2L types which possessed similar feature vectors to normal instances is obtained. Their study also demonstrated that a large training set is not necessary in obtaining high detection rates. However, the model lacked in detecting routine attack types. Synthetic minority oversampling technique (SMOTE) was used to balance the training sets at the beginning. Due to optimal dimension reduction, the computation time and complexity of their model were lessened overall.

Naseer et al. (2018) built a deep learning (DL)-based NIDS model which leveraged convolutional neural networks (CNNs), autoencoders (AEs) and LSTM networks. The most relevant features were extracted through a bottleneck layer, after training of the AEs using NSKLDDTrain+ dataset. The features selection phase which reduced dataset dimensionality from 41 to 16 was then fed to a multi-layer perceptron for anomaly detection. As part of their evaluation, they compared the performance of their models to that of conventional classification techniques such as K-NN, Random Forest (RF), NB and several others. The NSL-KDD dataset was used in the training stage, while the NSLKDDTest+ and NSLKDDTest21 were used in the validation stage, in all experiments. The DNN and LSTM models yielded 85% and 89% test accuracy, showing that DL has promising potential in building a high performing NIDS.

Chohra et al. (2022) built a NIDS in which its feature selection phase was optimised through a particle swarm optimisation (PSO) technique merged with ensemble methods. Two types of fitness functions were explored; bagging method and boosting method. It was found that more optimal solutions were found through boosting ensemble than bagging, which significantly reduced training time delays. The fitness function yielded values of Acc, Rec, Prec and F1, with the objective function defined as the maximisation of F1 to prevent the algorithm from falling into a local optimum. The extracted features were then sent to a DL-based AE anomaly detector, with tuned hyperparameters being the batch size, loss function, number of layers, number of neurons and regularisations. For evaluation, IoT-Zeek, NSL-KDD and UNSWNB15 datasets were used. The IoT-Zeek dataset was generated by the authors to demonstrate the model's applicability towards latest cyber-threat scenarios. The first set of models comprised four classical ML methods; RF, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting machine (lightGBM) and Category Boosting (CatBoost), the second set was two DL methods; CNN, Feed-Forward Neural Network (FFNN), and finally, ensemble learning was done by averaging the four classical and two DL models, with all models having equal contribution. The proposed AE showed higher F1 values for NSL-KDD and UNSWNB15 datasets.

Onah et al. (2021) proposed a NIDS model which was built on a genetic algorithm (GA) wrapper-based feature selection approach and NB classifier. The dataset used for evaluation was NSL-KDD. The dimension of the dataset was reduced in the feature selection stage, and classification of the reduced dataset was performed using NB for a NIDS in a fog computing environment. The performance metrics focused on were time, prec, and acc.

Zhong et al. (2020), whom evaluated their model on the IDS2017 dataset, built a NIDS model which was based on

heterogenous ensemble learning. Their research problems focused on firstly, the inadaptability of models in real world cyber-attacks due to training being limited to stale datasets; secondly, the inability of most algorithms in learning new models when the attack environment is varied; and thirdly, complex network attack environments which cannot be dealt by single detection algorithms which have only one peak value. Their methodology comprises using a Damped Incremental Statistics (DIS) algorithm for feature extraction, which reduces sample weight over time. Next, an AE was trained using some labelled data, and was used to mark the anomalous score of network traffic. Furthermore, data labelled as anomalous was used to train the LSTM. Finally, to calculate the final anomalous score, weighted averaging was done using the results obtained from the AE and LSTM. The Mirai dataset in Kitsune was used to observe the effects of LSTM predictions. A DBN used in the feature extraction stage reduced the dimensionality, improving the efficiency of the training AE. The anomalous scores of three samples were predicted using those of historical samples, and the LSTM was used to predict timing on anomalies. A discriminant formula is then applied to obtain the anomaly detection score for each packet, with a simulated annealing algorithm to optimally select values of *gp* (dynamically changing anomaly detection threshold) and *p* (coordinate the weight between precited and detected values). The MAWILab dataset in the June 3, 2018 period was used in their experiment, and the IDS2017 dataset was used for comparison with other algorithms. Their ensemble model yielded better performance metrics compared to single models. Table V shows a summary of the existing NIDS models previously discussed.

TABLE V. SUMMARY OF REVIEWED EXISTING NIDS MODELS

Dataset	Ref	Feature Selection	Classifier	Performance Metrics (%)				
				Acc	Prec	Re	F1	AuC
CIC-ID2017	(Ding and Li, 2022)	GCN, LSTM	SoftMax	99.24	98.50	98.62	98.72	99.0
	(Chen et al., 2022)	DBN	LSTM	Normal – 99.84 DoS – 99.72 Portscan – 99.79 Brute Force – 82.16 Web Attack – 50.24	-	-	-	-
	(Aksu and Aydin, 2022)	MGA m-feature selection	MGA-DTC	100.0	100.0	100.0	100.0	100.0
	(Panigrahi et al., 2022)	MOEFS	DTNB	96.5	97.4	96.7	-	-
NSL-KDD	(Pajouh et. al, 2017)	LDA	NB, K-NN-CF	-	-	Normal – 94.56 Probe – 79.76 DoS – 84.68 U2R – 67.16 R2L – 34.81	-	-
	(Naseer et al., 2018)	DIS, DBN, AE	LSTM, Weighted averaging	NSLKDDTest+ – 89.0 NSLKDDTest21 – 83.0	-	-	-	NSLKDDPlus – 95.50 NSLKDD21 – 91.6
	(Chohra et al., 2022)	PSO + 6 ensemble models	AE	90.71	89.351	95.005	92.092	95.4
	(Onah et al., 2021)	Wrapper-based GA	NB	99.73	99.10	-	-	-

IV. CONCLUSION AND FUTURE RESEARCH

Due to the large dimensionality and complexity of the cybersecurity datasets, appropriate measures must be taken to ease decision-making, as well as to reduce the complexity and dimensionality of the dataset. Therefore, feature selection is an essential stage of a NIDS, as only the most relevant features will be extracted from a given dataset and fed into the classifier, ultimately improving detection rate as well as model accuracy.

Machine learning models highly rely on dataset features, sample distribution and the sample amount, which limits the extent of their performance to datasets which are uniformly distributed, have adequate samples and well-selected features. Therefore, to address real-time cyber threats, a NIDS would require orderly updates. This leads to future research which should focus on more effective feature extraction techniques and enhancement of scalability towards the emergence of unseen attacks. Furthermore, based on the recent work discussed, deep learning methods and hybrid techniques show great promise in the development of robust NIDS for future research.

REFERENCES

Ahmed, M., Mahmood, A., Islam, M. (2016) A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems, 55,* 278-288, https://doi.org/10.1016/j.future.2015.01.001

Aksu, D., & Aydin, M. A. (2022). MGA-IDS: Optimal feature subset selection for anomaly detection framework on in-vehicle networks-CAN bus based on genetic algorithm and intrusion

detection approach. *Computers & Security*, *118*, 102717. https://doi.org/10.1016/j.cose.2022.102717

Canadian Institute for Cybersecurity. (1999). *NSL-KDD Dataset.* https://www.unb.ca/cic/datasets/nsl.html

Canadian Institute for Cybersecurity. (2017). *Intrusion Detection Evaluation Dataset (CIC- IDS2017).* https://www.unb.ca/cic/datasets/ids-2017.html

Chen, A., Fu, Y., Zheng, X., & Lu, G. (2022). An efficient network behavior anomaly detection using a hybrid DBN-LSTM network. *Computers & Security*, *114*, 102600. https://doi.org/10.1016/j.cose.2021.102600

Chohra, A., Shirani, P., Karbab, E. B., & Debbabi, M. (2022). Chameleon: Optimized feature selection using particle swarm optimization and ensemble methods for network anomaly detection. *Computers & Security*, *117*, 102684. https://doi.org/10.1016/j.cose.2022.102684

Ding, Q., & Li, J. (2022). AnoGLA: An efficient scheme to improve network anomaly detection. *Journal of Information Security and Applications*, *66*, 103149. https://doi.org/10.1016/j.jisa.2022.103149

Hossin, M., & M.N, S. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5(2)*, 3-5. https://doi.org/10.5121/ijdkp.2015.5201

Naseer, S., Saleem, Y., Khalid, S., Bashir, M. K., Han, J., Iqbal, M. M., & Han, K. (2018). Enhanced Network Anomaly Detection Based on Deep Neural Networks. *IEEE Access*, *6*, 48231-48246. doi: 10.1109/ACCESS.2018.2863036

National Technology Security Coalition (2020). *Cyber Security Report 2020. https://www.ntsc.org/assets/pdfs/cyber-security-report-2020.pdf*

Onah, J. O., Abdulhamid, S. i. M., Abdullahi, M., Hassan, I. H., & Al-Ghusham, A. (2021). Genetic Algorithm based feature selection and Naïve Bayes for anomaly detection in fog computing environment. *Machine Learning with Applications*, *6*, 100156. https://doi.org/10.1016/j.mlwa.2021.100156

Pajouh, H. H., Dastghaibyfard, G., & Hashemi, S. (2017). Two-tier network anomaly detection model: a machine learning approach. *Journal of Intelligent Information Systems*, *48*(1), 61-74. https://doi.org/10.1007/s10844-015-0388-x

Panigrahi, R., Borah, S., Pramanik, M., Bhoi, A. K., Barsocchi, P., Nayak, S. R., & Alnumay, W. (2022). Intrusion detection in cyber–physical environment using hybrid Naïve Bayes—Decision table and multi-objective evolutionary feature selection. *Computer Communications*, *188*, 133-144. https://doi.org/10.1016/j.comcom.2022.03.009

Scarfone, K., & Mell, P. (2010). Intrusion Detection and Prevention Systems. In P. Stavroulakis & M. Stamp (Eds.), *Handbook of Information and Communication Security*, 177-192. https://doi.org/10.1007/978-3-642-04117-4_9

Sharafaldin, I., Gharib, A., Habibi Lashkari, A., & Ghorbani, A. (2017). Towards a Reliable Intrusion Detection Benchmark Dataset. *Software Networking*, *2017(1)*, 177-200. https://doi.org/10.13052/jsn2445-9739.2017.009

Singh, G., & Khare, N. (2021). A survey of intrusion detection from the perspective of intrusion datasets and machine learning techniques. *International Journal of Computers and Applications*, 1-11. https://doi.org/10.1080/1206212X.2021.1885150

Zhong, Y., Chen, W., Wang, Z., Chen, Y., Wang, K., Li, Y., . . . Li, K. (2020). HELAD: A novel network anomaly detection model based on heterogeneous ensemble learning. *Computer Networks*, *169*, 107049. https://doi.org/10.1016/j.comnet.2019.107049