Dimensionality Reduction of Network Intrusion Detection System Using Machine Learning

SHAFIN-UL-ALAM, Bangladesh University of Engineering and Technology, Bangladesh
M SHAHIR RAHMAN, Bangladesh University of Engineering and Technology, Bangladesh
A S M NAZRUL ISLAM, Bangladesh University of Engineering and Technology, Bangladesh

Abstract- Growing attacks in the domain of the Internet have guided the demand for intrusion detection systems (IDS). As a result, numerous researchers have tried to improve IDS's performance. A network-based IDS's core idea is to scan network packets at the router or host level and audit packet information, logging any suspicious packets into a distinct log file with extended information. A number of published studies show that NSL-KDD dataset is widely claimed as a benchmark intrusion detection dataset as it does not include redundant records. In this research, we have lowered the dimensionality of the dataset to reduce computation time and ensure improved performance in real-time since high amounts of features might lead to machine learning algorithms performing poorly. Without substantially losing any information, a significant number of features are removed. However, it has also come with a cost of accuracy. There are several approaches to determine whether the feature reduction process is reliable or not. In order to observe the outcomes, various algorithms are applied on the dataset which provides us with distinct results. The article also includes these findings.

Additional Key Words and Phrases: IDS, Network, Intrusion Detection, Dimensionality Reduction

1 Introduction

Technology advancements increase the risk of a computer's security. With the rapid development of the Internet, the issue of cyber security has increasingly gained more attention. An intrusion Detection System (IDS) is an effective technique to defend against cyber-attacks and reduce security losses. The purpose of IDS is to help computer systems deal with attacks. This anomaly detection system makes a database of expected behavior and deviations from the expected behavior to trigger during intrusions. To acquire a high level of threat visibility, organizations must ensure that intrusion detection technology is correctly installed and optimized. Because of budget and monitoring constraints, it may not be rational to place HIDS and NIDS sensors throughout an IT environment. Investigating IDS notifications can be very resource and time-intensive, requiring extra information from other systems to help fix whether an alarm is serious. A generic problem for organizations that implement IDS is that they seldom have an appropriate incident response capability.

To achieve a top level of threat visibility, organizations must make sure that intrusion detection system is properly optimized. Many organizations lack a complete overview of their network, implementing IDS can be tricky and, if not done correctly, they may leave critical assets exposed. Signature and anomaly-based detection techniques are often used in HIDS and NIDS. This implies that when a sensor detects activity that matches a known attack pattern or traffic that deviates from a set of standard behaviors, an alert is issued. High-bandwidth usage and erratic web or DNS traffic might be signs of unusual behavior.

Intrusion detection generates a large number of notifications, which may be a substantial load for internal teams. Many system alarms are false positives, but businesses seldom have the time or resources to thoroughly investigate every signal, allowing suspicious activity to go undetected. Most intrusion detection systems come with a collection of pre-defined alarm signatures, but they are insufficient for most businesses, necessitating further work to baseline behaviors particular to each environment. IDS warnings include basic security information that, when seen in isolation, may be meaningless. When you receive an alert, it's not always clear what generated it or what activities you need to do to determine whether or not it's a legitimate threat.

Concentrating on the field of interruption identification initially began in 1980 and the principal such model was distributed in 1987 [1]. Throughout the previous few decades, however tremendous business speculations and considerable exploration were done, interruption identification innovation is as yet juvenile and subsequently not compelling [1]. While network IDS that works dependent on signature have seen business achievement and far reaching reception by the innovation based association all through the globe, inconsistency based organization IDS have not acquired accomplishment in a similar scale. Because of that explanation in the field of IDS, at present inconsistency based identification is a significant center space of innovative work [2]. What's more, prior to going to any wide scale organization of inconsistency based interruption identification framework, central points of contention still need to be addressed [2]. Be that as it may, the writing today is restricted with regards to think about on how interruption identification performs when utilizing administered AI procedures [3].

To ensure target frameworks and organizations against malevolent exercises inconsistency based organization IDS is a significant innovation. Notwithstanding the assortment of inconsistency based organization interruption identification strategies portrayed in the writing as of late [2], irregularity recognition functionalities empowered security apparatuses are simply starting to show up, and some significant issues still need to be addressed. A few inconsistency based strategies have been proposed including Linear Regression, Support Vector Machines (SVM), Genetic Algorithm, Gaussian combination model, knearest neighbor calculation, Naive Bayes classifier, Decision Tree [4]. Among them the most generally utilized learning calculation is SVM as it has effectively laid down a good foundation for itself on various sorts of issue [5]. One significant issue on inconsistency based identification is however

Authors' addresses: Shafin-Ul-Alam, Bangladesh University of Engineering and Technology, Bangladesh, shafinmist@gmail.com; M Shahir Rahman, Bangladesh University of Engineering and Technology, Bangladesh, shahir.rahman3502@gmail.com; A S M Nazrul Islam, Bangladesh University of Engineering and Technology, Bangladesh, nazrul.islam@biman.gov.bd.

this large number of proposed procedures can distinguish novel assaults yet they all experience a high bogus caution rate overall. The reason behind is the intricacy of producing profiles of down to earth ordinary conduct by gaining from the preparation informational collections [6]. Today Artificial Neural Network (ANN) are frequently prepared by the back proliferation calculation, which had been around beginning around 1970 as the converse method of programmed separation [7]. The significant difficulties in assessing execution of organization IDS is the inaccessibility of a far reaching network based informational collection [8]. The vast majority of the proposed inconsistency based strategies found in the writing were assessed utilizing KDD CUP 99 dataset [9].

In this paper, our *objective* is to apply different machine learning algorithms to evaluate the overall effectiveness of NIDS after using various dimensionality reduction techniques to remove insignificant features while retaining the most relevant ones. We have also shown a comprehensive evaluation of the results regarding the correctness of each procedure after reducing the dimension of the dataset to demonstrate the efficacy of the dimensionality reduction method and the performace of NIDS.

Our *contributions* of this work are: (i) Reducing the dimensionality of a large dataset using three different feature reduction technique. (ii) Lowering the number of features without compromising the overall performance of NIDS. (iii) Comparative of study regarding the efficiency of dimension reduction techniques by demonstrating each process' correctness after applying several machine learning algorithms. After reducing the dimensionality using three distinct methodologies, we applied ANN, SVM, LSTM, and MLP machine learning algorithms to the NSL-KDD dataset.

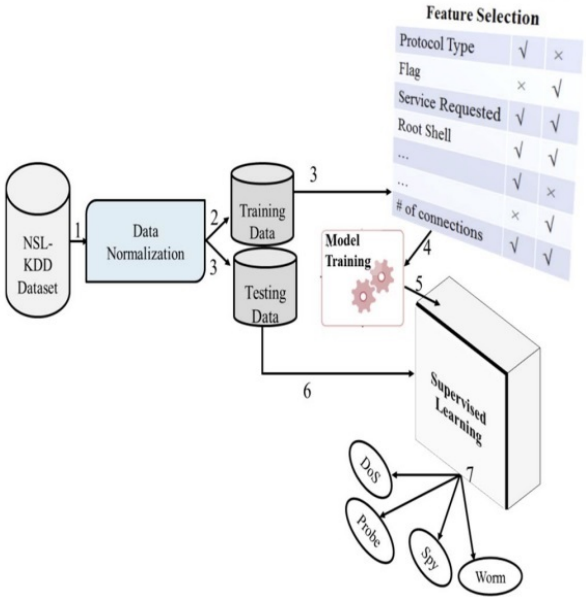
This paper is coordinated as follows: The most recent contributions to this field are summarized in Section 2. In section 3 we have presented our methodology along with the flow diagrams and appropriate description. The outcome of the total process is container in Section 4. Finally, we conclude the paper in Section 5.

2 Related Works

2.1 Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection by Kazi Abu Taher et. al Their proposed modeltried to find the best classifier with higher accuracy and success rate. The system proposed is made of feature selection and supervising algorithm show in Fig. Feature selection component are responsible to extract most relevant features or attributes to identify the instance to a specific group or class. The learning algorithm component builds the important intelligence or knowledge using the result found from the feature selection component. Using the training dataset, the model gets trained and builds its intelligence. Then the learned intelligences are applied to the testing dataset to measure the accuracy of home much the model correctly classified on unseen data.

For highlight determination channel technique and covering strategy have been utilized. In channel strategy, highlights are chosen on the premise of their scores in different factual tests that action the significance of highlights by their connection with subordinate variable or result variable. Covering strategy sees as a subset of elements by estimating the handiness of a subset of highlight with the reliant variable. Henceforth channel techniques are autonomous of any AI calculation while in covering strategy the best component subset chose depends on the AI calculation used to prepare the model. In covering strategy a subset evaluator utilizes all conceivable subsets and afterward utilizes an order calculation to persuade classifiers from the elements in every subset. The classifier consider the subset of element with which the grouping calculation plays out great.

In this paper, they have introduced distinctive machine learning models utilizing distinctive AI calculations and diverse component determination techniques to track down a best model. The investigation of the outcome shows that the model constructed utilizing ANN and covering highlight determination outflanked any remaining models in characterizing network traffic effectively with identification pace of 94.02 percent.



2.2 User behavior Pattern -Signature based Intrusion Detection by Zakiyabanu S. Malek et. al To recognize the ordinary and strange client conduct we distinguish client boundaries and develop the client's underlying conduct ordinary/regular profile. This profile isn't consistent and may fluctuate according to the utilization of the framework (Change in client conduct). In their [10] standard based concentrate on it includes rules and the principles are either planned by the framework or given by a specialist. These principles are applied to the accumulated client conduct subtleties. The principles are put away in a standard motor. Client conduct subtleties comprise of client conduct log while rules involve as an on the off chance that explanation which is simple for individuals to comprehend by applying these guidelines on client conduct subtleties it can without much of a stretch distinguish where it is meddlesome or not. On the off chance that the standard is set off, then, at that point, the framework reports to obstruct a record of the client or may erect a caution to inform.

They have utilized the client profile dataset of Zakiya [11] which contains boundaries distinctive of a console, Mouse, applications

running, processor utilization, and so forth Zakiya had fostered a measurable motor which apply calculated relapse and Statistical mean on distinctive client's dataset and experiments with various highlights. To proceed with the work we relegate an master. Presently, specialists knowing ordinary client conduct subsequently, the master gives rules in Pattern based Interruption Detection Engine PIDE. Here, we have utilized JESS to give rules. The standard based identification gathers different information for potential assaults to distinguish approved furthermore, unapproved exercises. Rules are to be characterized in such a way that main far fetched exercises are taken note without upsetting approved clients.

2.3 An intrusion detection system integrating networklevel intrusion detection and host-level intrusion detection by Jiannan Liu et. al They propose an original NN-based half and half IDS structure, AlarmNet-IDS. NIDS and HIDS cover various scopes of assaults and the blend makes the scope of identification bigger. For those assaults in the two territories, this technique can work on the exactness and unwavering quality of identification. The proposed structure was carried on ARMv8-based implanted stage. Reproduction of a few assaults in the genuine climate and accomplish great outcomes. It is a productive IDS structure which can precisely respond to the interruption on schedule. In the plan of NIDS, an autoencoder is applied as a component student in NN-based choice motor. The autoencoder can give more discriminative highlights. We use it notwithstanding the first highlights to work on the presentation of the HIDS model. For the HIDS model, we gather framework call hints of each interaction as crude information. Our methodology consolidated word installing strategies and convolutional neural organization to deal with these information. Word installing helps map framework calls to highlight vectors to catch the semantic comparability and the convolutional neural organization, great at portrayal learning helps extricate the elements of framework call follows.

On NSL-KDD, they [12] developed their NIDS models with a three-layer autoencoder and a three-layer completely associated neural organization where they change the number of neurons in the center layers of them individually. They pick 32 and 64 for the autoencoder and they look over 32 to 512 for the completely associated neural organization. Autoencoders as inactive highlights are an interesting technique and on the off chance that it acts well, they will add it, while if not, they can eliminate it. Additionally, they perform other AI calculations like SVM, LR, KNN, NB as a differentiation. They ascertain exactness, accuracy, review, and F1 score for both different characterization and parallel arrangement. For parallel characterization, they additionally ascertain AUC by doing ROC bends. They find that the model with 64 neurons in the autoencoder and 256 neurons in the neural organization gets the most noteworthy AUC in their investigations. On ADFALD, we play out their HIDS model and they utilize various sizes of convolutional portions from 2 to 6. Additionally, they perform different calculations like LR, SVM, KNN, NB

They propose a NN-based half and half structure, AlarmNet-IDS, for interruption identification, which consolidates NIDS and HIDS. In the NIDS model, they utilized autoencoders for programmed highlight extraction and neural organizations for characterization. For HIDS, they use word installing techniques from NLP that maps framework calls into highlight vectors and afterward send them to one-dimensional convolutional neural organizations to extricate transient elements for characterization. They perform investigates public benchmarks, NSL-KDD and ADFA-LD and their models accomplish great execution. They convey our IDS structure on the Linux-based implanted terminal and confirm it in the genuine climate.

2.4 Analysis of Heuristic based Feature Reduction method in Intrusion Detection System. by Swapnil Umbarkar et. al Notwithstanding of having bunches of exploration on highlight decrease, there is no specific investigation made on reduction strategies like which procedure is better than different methods and in regard to what. Their work investigates IG, GR and Correlation to downsize the quantity of highlights without corrupting the presentation of the framework and furthermore exhibit the idea of each component decrease procedure concerning number of elements lastly gives the best elements decrease strategy from previously mentioned method with least number of elements.

2.5 Intrusion Detection System Using Feature Selection and Classification Technique by Senthilnayaki Balakrishnan et al With the quick development of web correspondence and accessibility of instruments to interrupt the organization, security for network has become indispensable.security arrangements don't adequately monitor the information put away in the data sets. To shield frameworks from being assaulted by gatecrashers, another Intrusion Detection System has been proposed and carried out in this undertaking work, which consolidates a straightforward component determination calculation and SVM procedure to distinguish assaults. Utilization of KDD cup informational collection and Data Mining extricate the concealed prescient data from enormous Databases which is an incredible new innovation with extraordinary potential. Insightful Agent based Attribute Selection Algorithm has been created by ascertaining Data Gain Ratio. They diminished arrangement of highlights R from set of 41 elements from KDD'99 Cup informational collection. Restrictions are that the KDD'99 cup informational collection is seen to be enormous and the proposed highlight determination calculation chooses just the significant highlights.

In their work, another IDS has been proposed and carried out by consolidating an Optimal Feature Selection (OFS) calculation and two characterization procedures for getting the framework. The calculation time taken for distinguishing and characterizing the records utilizing every one of the 41 highlights of the KDD'99 cup informational collection is seen to be enormous. The proposed highlight determination calculation chooses just the significant elements that assistance in diminishing the time taken for distinguishing and characterizing the records. Further the standard based classifier and SVM assist with accomplishing a more noteworthy exactness. The fundamental benefit of the proposed IDS is that it diminishes the bogus positive rates and furthermore decreases the calculation time.

3 Problem Formulation and Methodology

3.1 Dataset There are many public datasets available for network intrusion detection. We selected NSL-KDD dataset for our experiment. It is a public dataset. NSL-KDD data set is a refined version of its predecessor KDD99 data set [13]. The number of records in the NSL-KDD sets are reasonable. So there is no need to run a random small portion. Researchers can run the full dataset. It is the main advantage of this dataset.

	KDD train set	KDD test set
Attacks	262,178	29,378
Normal	812,814	47,911
Total	1,074,992	77,289

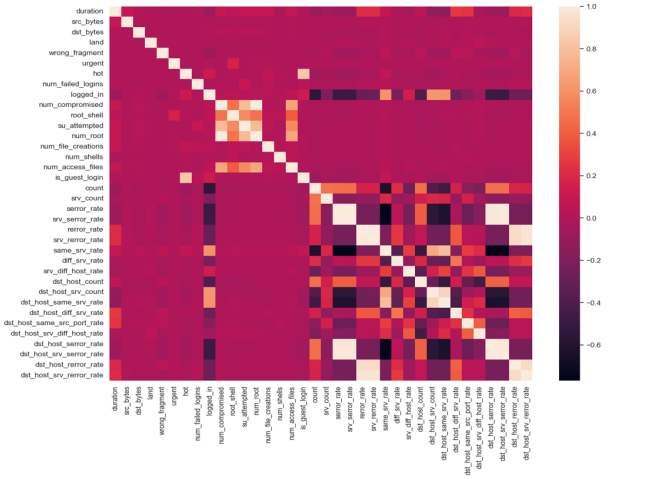
There are total 1,074,992 records on the training dataset and 77,289 records on the test dataset. In the NSL-KDD total 42 features are available including 1 class label. Most of the features are with numeric values. But three features: Protocol type, Service, Flag are categorical features.

F#	Feature name	F#	Feature name	F#	Feature name
F1	Duration	F15	Su attempted	F29	Same srv rate
F2	Protocol type	F16	Num root	F30	Diff srv rate
F3	Service	F17	Num file creations	F31	Srv diff host rate
F4	Flag	F18	Num shells	F32	Dst host count
F5	Source bytes	F19	Num access files	F33	Dst host srv count
F6	Destination bytes	F20	Num outbound cmds	F34	Dst host same srv rate
F7	Land	F21	Is host login	F35	Dst host diff srv rate
F8	Wrong fragment	F22	Is guest login	F36	Dst host same src port rate
F9	Urgent	F23	Count	F37	Dst host srv diff host rate
F10	Hot	F24	Srv count	F38	Dst host serror rate
F11	Number failed logins	F25	Serror rate	F39	Dst host srv serror rate
F12	Logged in	F26	Srv serror rate	F40	Dst host rerror rate
F13	Num compromised	F27	Rerror rate	F41	Dst host srv rerror rate
F14	Root shell	F28	Srv rerror rate	F42	Class label

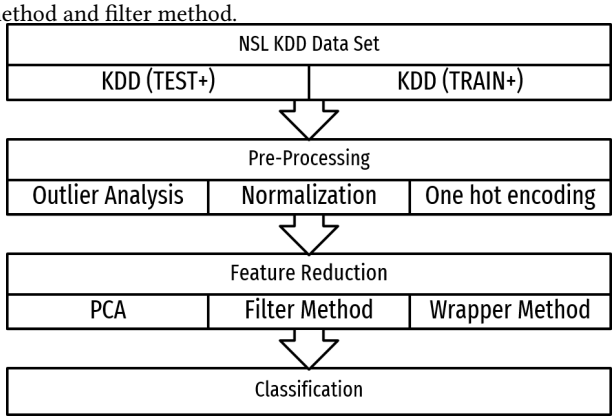
In the class label total 37 types of attacks are introduced with 5 attack class. Attack classes are: Dos, Probe, R2L, U2R.

Attack Class	Attack Type
DoS	Back, Land, Neptune, Pod, Smurf,Teardrop,Apache2, Udpstorm, Processtable, Worm (10)
Probe	Satan, Ipsweep, Nmap, Portsweep, Mscan, Saint (6)
R2L	Guess_Password, Ftp_write, Imap, Phf, Multihop, Warezmaster, Warezclient, Spy, Xlock, Xsnoop, Snmpguess, Snmpgetattack, Httptunnel, Sendmail, Named (16)
U2R	Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps (7)

3.2 Preprocessing Normalization is a part of data preprocessing in machine learning. The process of normalization is to convert the values of numeric features in dataset to a common scale without losing any information. Some algorithms requires normalization to model the data properly. In our experiment we used normalization in our preprocessing step so that we can model the data more correctly. We have total 42 features in our dataset. Out of 42 features 39 features have numeric value and 3 features are categorical. Some machine learning algorithms cannot operate on categorical data directly. So it is necessary to covert the categorical features into numeric features. One-Hot encoding is a technique which is used to convert categorical data to integer data. So we used One-Hot encoding in out preprocessing step. The three categorical features protocol type, service, flag are converted into numerical values with the help of one-hot-encoding technique.



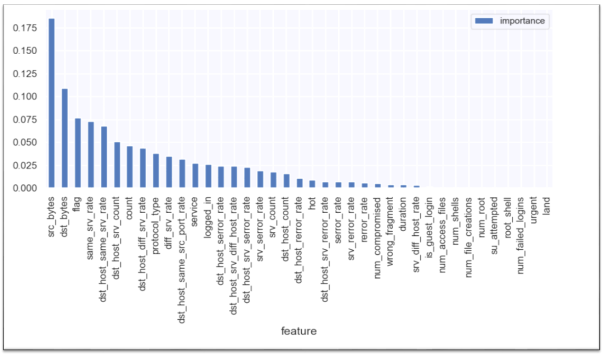
3.3 Feature Reduction Feature reduction or dimensionality reduction is the process of reducing the number of features without losing much information . Reduced number of features makes the machine learning model faster and easier. Besides that feature reduction makes the data visualization easier for humans. In this paper we used three dimensionality reduction techniques: PCA, Wrapper method and filter method.



3.3.1 PCA PCA is used for dimensionality reduction in machine learning. It transforms a large set of correlated features into a set of linearly uncorrelated features that still contains most of the information. These new transformed features are called Principal Components. In our experiment we transformed all the features of our dataset into 10 principal components. Reducing the number of variables of a data set most of the time comes at the expense of accuracy.

3.3.2 Wrapper Method Wrapper method is another dimensionality reduction method which is based on a specific machine learning algorithm that we try to fit on a given dataset. Feature selection can be forward selection and backward elimination. In forward selection we start with 1 feature and each step we add a new feature. But in backward elimination we start with all the features and each step we eliminate a feature. In our experiment after applying wrapper method we got 20 features. Then we conducted all the all the models with this 20 features and stored the experimental results for comparison.

3.3.3 Filter Method Filter method is type of dimensionality reduction method where selection of features is independent of any machine learning algorithms. Features are selected on the basis of their scores in various statistical tests. We used Pearson correlation coefficient statistical test to score the features. Finally we got 19 features in this method. We conducted all the all the models with this 19 features and stored the experimental results for comparison.



In the above figure, we have plotted the features according to importance. It is made apparent that source bytes, destination bytes, flag, as well as other similar features are plotted as being crucial for NIDS characteristics since these features are necessary to identify an attack by examining certain patterns in source or destination bytes or flag values.

3.3.4 Classification In our experiment we used both shallow learning and deep learning algorithms to detect the network intrusion. We used total 4 algorithms:

- Support Vector Machine (SVM)
- Artificial Neural Network (ANN)
- Long Short-Term Memory (LSTM)
- Multilayer Perceptron (MLP)

Details demonstration of classification result are shown in the next section.

4 Experimental Results

We have followed the methodology to reduce the dimensionality and afterwards to ensure the efficacy of the feature reduction methods, we applied four different ML algorithms. The results are presented below:

Methods	Accuracy with 41 features	PCA	Wrapper	Filter
ANN	0.8092	0.7943	0.7857	0.7979
SVM	0.7728	0.7597	0.7761	0.7887
LSTM	0.8009	0.8066	0.7972	0.8200
MLP	0.8112	0.7891	0.7993	0.7957

We used four deep learning and shallow learning algorithms with three dimensionality reduction method. For each algorithm we get:

- Accuracy with all features
- Accuracy with PCA
- Accuracy with Wrapper Method
- Accuracy with Filter Method

We get the accuracy score of three dimensionality reduction method. We can see from the table, Filter method has higher accuracy score than other two methods. We also see that out of four algorithms we get higher accuracy with LSTM model which is a deep learning model. For our experiment deep learning models are more suitable than shallow learning models.

We also demonstrated the detection time of different models with different dimensionality reduction method. We can see from the table that SVM takes highest intrusion detection time. SVM is a shallow learning method. So for our experiment shallow learning algorithms has higher detection time. If we compare the detection time with different dimensionality methods we can see from the table, Wrapper method has the lowest detection time for most of the models.

Method	Detection time with 41 features	PCA	Wrapper	filter
ANN	1.3488	0.6274	0.4631	0.6917
SVM	8.1922	8.6910	7.4851	8.5672
LSTM	0.8427	0.8267	0.7497	0.8370
MLP	0.8619	0.6725	0.6995	0.6471

5 Conclusion

In this paper, we have applied three potential dimensionality reduction methods to reduce the features from a large dataset with a view to improving the performance and computational speed of Network Intrution Detection system. Afterwards, in order to ensure the efficacy of the aforementioned dimensionality reduction methods, we applied different machine learning algorithms in order to assess the overall performance of the system. Moreover, we have shown a comparative analysis of accuracy between these results obtained from the ML algorithms to comprehend the efficacy and gain a thorough understanding of the entire procedure. Comparing

the Filter Method to PCA and Wrapper Method, we have observed that the Filter method performed better in this trial with more accuracy. We also noted that deep learning models outperform other models in terms of accuracy. We firmly believe that the results of our research will enable the researchers to make a significant contribution to this area of network intrusion detection systems in future.

References

[1] N. Chakraborty, "Intrusion detection system and intrusion prevention system: A comparative study," *International Journal of Computing and Business Research (IJCBR)*, vol. 4, no. 2, pp. 1–8, 2013.

[2] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *computers & security*, vol. 28, no. 1-2, pp. 18–28, 2009.

[3] M. C. Belavagi and B. Muniyal, "Performance evaluation of supervised machine learning algorithms for intrusion detection," *Procedia Computer Science*, vol. 89, pp. 117–123, 2016.

[4] M. Saber, S. Chadli, M. Emharraf, and I. El Farissi, "Modeling and implementation approach to evaluate the intrusion detection system," in *International conference on networked systems*, pp. 513–517, Springer, 2015.

[5] J. Zheng, F. Shen, H. Fan, and J. Zhao, "An online incremental learning support vector machine for large-scale data," *Neural Computing and Applications*, vol. 22, no. 5, pp. 1023–1035, 2013.

[6] F. Gharibian and A. A. Ghorbani, "Comparative study of supervised machine learning techniques for intrusion detection," in *Fifth Annual Conference on Communication Networks and Services Research (CNSR'07)*, pp. 350–358, IEEE, 2007.

[7] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[8] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 military communications and information systems conference (MilCIS)*, pp. 1–6, IEEE, 2015.

[9] T. Janarthanan and S. Zargari, "Feature selection in unsw-nb15 and kddcup'99 datasets," in *2017 IEEE 26th international symposium on industrial electronics (ISIE)*, pp. 1881–1886, IEEE, 2017.

[10] L. Dhanabal and S. Shantharajah, "A study on nsl-kdd dataset for intrusion detection system based on classification algorithms," *International journal of advanced research in computer and communication engineering*, vol. 4, no. 6, pp. 446–452, 2015.

[11] D. Hu, "An introductory survey on attention mechanisms in nlp problems," in *Proceedings of SAI Intelligent Systems Conference*, pp. 432–448, Springer, 2019.

[12] J. Liu, K. Xiao, L. Luo, Y. Li, and L. Chen, "An intrusion detection system integrating network-level intrusion detection and host-level intrusion detection," in *2020 IEEE 20th International Conference on Software Quality, Reliability and Security (QRS)*, pp. 122–129, IEEE, 2020.

[13] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *2009 IEEE symposium on computational intelligence for security and defense applications*, pp. 1–6, IEEE, 2009.