International Journal of Scientific Research in Science and Technology
Available online at : www.ijsrst.com
OPEN ACCESS
Print ISSN: 2395-6011 | Online ISSN: 2395-602X
doi : https://doi.org/10.32628/IJSRST251222614

# Exploring the NSL-KDD Dataset: A Comprehensive Analysis about Intrusion Detection System (IDS)

**Bechoo Lal*, Thoudam Basanta Singh[2], Mutum Bidyarani Devi[3]**

*Research Scholar, Department of Computer Science and Engineering, Manipur International University (MIU), Imphal West– 795140, Manipur, India

[2]Department Department of Physics, Position: Professor and Dean of the School of Physical Sciences and Engineering, Manipur International University, Manipur, India
Orchid : 0009-0007-4254-5286

[2]Department of Electronics and Computer Sciences, School of Physical Sciences and Engineering, Assistant Professor, Manipur International University, Manipur, India
Orchid : 0000-0001-7270-8620

ARTICLEINFO

**Article History:**

Accepted : 05 April 2025
Published: 08 April 2025

**Publication Issue :**
Volume 12, Issue 2
March-April-2025

**Page Number :**
715-722

ABSTRACT

In this research article the researcher emphasized the Network threats and hazards are evolving at a high-speed rate in recent years. Many mechanisms (such as firewalls, anti-virus, anti-malware, and spam filters) are being used as security tools to protect networks. An intrusion detection system (IDS) is also an effective and powerful network security system to detect unauthorized and abnormal network traffic flow. This article presents a review of the research trends in network-based intrusion detection systems (NIDS), their approaches, and the most common datasets used to evaluate IDS Models. The analysis reported presented in this paper is based on the supervised machine learning approach logistics and XGB-classifier by using NSL-KDD Dataset. The researcher found that logistic classifier given 0.95% accuracy where as XGBooster Classifier gives the 1.00% accuracy , due to the over fitting the researcher used the hyper parameter tuning XGB classifier and got the 0.99% accuracy. The researcher assured that the developed predictive model is more accurate and efficient to detect the intrusion during the data transmission.

**Keywords:** IDS, Network Security, NSL-KDD, Machine Learning, XGB Classifier, Logistic Regression

## I. INTRODUCTION

The evolution of malicious software (malware) poses a critical challenge to the design of intrusion detection systems (IDS). Malicious attacks have become more sophisticated and the foremost challenge is to identify unknown and obfuscated malware, as the malware authors use different evasion techniques for information concealing preventing detection by an IDS. In addition, there has been an increase in

security threats such as zero-day attacks designed to target internet users. Therefore, computer security has become essential as the use of information technology has become part of our daily lives. As a result, various countries such as Australia and the US have been significantly impacted by the zero-day attacks. According to the 2017 Symantec Internet Security Threat Report, more than three billion zero-day attacks were reported in 2016, and the volume and intensity of the zero-day attacks were substantially greater than previously (Symantec, 2017).

As highlighted in the Data Breach Statistics in 2017, approximately nine billion data records were lost or stolen by hackers since 2013 (Breach_LeveL_Index, 2017). A Symantec report found that the number of security breach incidents is on the rise. In the past, cybercriminals primarily focused on bank customers, robbing bank accounts or stealing credit cards (Symantec, 2017). However, the new generation of malware has become more ambitious and is targeting the banks themselves, sometimes trying to take millions of dollars in one attack (Symantec, 2017). For that reason, the detection of zero-day attacks has become the highest priority.

High profile incidents of cybercrime have demonstrated the ease with which cyber threats can spread internationally, as a simple compromise can disrupt a business' essential services or facilities. There are a large number of cybercriminals around the world motivated to steal information, illegitimately receive revenues, and find new targets. Malware is intentionally created to compromise computer systems and take advantage of any weakness in intrusion detection systems. In 2017, the Australian Cyber Security Centre (ACSC) critically examined the different levels of sophistication employed by the attackers (Australian, 2017). So there is a need to develop efficient IDS to detect novel, sophisticated malware..

## II. LITERATURE REVIEW

In the last few decades, machine learning has been used to improve intrusion detection, and currently there is a need for an up-to-date, thorough taxonomy and survey of this recent work. There are a large number of related studies using either the KDD-Cup 99 or DARPA 1999 dataset to validate the development of IDSs; however there is no clear answer to the question of which data mining techniques are more effective. Secondly, the time taken for building IDS is not considered in the evaluation of some IDSs techniques, despite being a critical factor for the effectiveness of 'on-line' IDSs.

This paper provides an up to date taxonomy, together with a review of the significant research works on IDSs up to the present time; and a classification of the proposed systems according to the taxonomy. It provides a structured and comprehensive overview of the existing IDSs so that a researcher can become quickly familiar with the key aspects of anomaly detection. This paper also provides a survey of data-mining techniques applied to design intrusion detection systems. The signature-based and anomaly-based methods (i.e., SIDS and AIDS) are described, along with several techniques used in each method.

The complexity of different AIDS methods and their evaluation techniques are discussed, followed by a set of suggestions identifying the best methods, depending on the nature of the intrusion. Challenges for the current IDSs are also discussed. Compared to previous survey publications (Patel et al., 2013; Liao et al., 2013a), this paper presents a discussion on IDS dataset problems which are of main concern to the research community in the area of network intrusion detection systems (NIDS). Prior studies such as (Sadotra & Sharma, 2016; Buczak & Guven, 2016) have not completely reviewed IDSs in term of the datasets, challenges and techniques. In this paper, we provide a structured and contemporary, wide-ranging study on intrusion detection system in terms of

techniques and datasets; and also highlight challenges of the techniques and then make recommendations.

During the last few years, a number of surveys on intrusion detection have been published. Table 1 shows the IDS techniques and datasets covered by this survey and previous survey papers. The survey on intrusion detection system and taxonomy by Axelsson (Axelsson, 2000) classified intrusion detection systems based on the detection methods. The highly cited survey by Debar et al. (Debar et al., 2000) surveyed detection methods based on the behaviour and knowledge profiles of the attacks. Taxonomy of intrusion systems by Liao et al. (Liao et al., 2013a), has presented a classification of five subclasses with an in-depth perspective on their characteristics: Statistics-based, Pattern-based, Rule-based, State-based and Heuristic-based. On the other hand, our work focuses on the signature detection principle, anomaly detection, taxonomy and datasets.

Intrusion can be defined as any kind of un-authorized activities that cause damage to an information system. This means any attack that could pose a possible threat to the information confidentiality, integrity or availability will be considered an intrusion. For example, activities that would make the computer services unresponsive to legitimate users are considered an intrusion. An IDS is a software or hardware system that identifies malicious actions on computer systems in order to allow for system security to be maintained (Liao et al., 2013a). The goal of an IDS is to identify different kinds of malicious network traffic and computer usage, which cannot be identified by a traditional firewall. This is vital to achieving high protection against actions that compromise the availability, integrity, or confidentiality of computer systems. IDS systems can be broadly categorized into two groups: Signature-based Intrusion Detection System (SIDS) and Anomaly-based Intrusion Detection System (AIDS).

Enterprises, governments, and people who depend on digital platforms need robust network security software. To safeguard a network and its resources from threats like viruses and illegal access, a reliable and durable security solution is necessary. The software and hardware components needed for network security protocols are listed by John Borky and Thomas Bradley (2019) as firewalls, routers, and anti-malware programs. Network security language is vital for providing dependable access, enhancing network speed, protecting all clients and data, and designing protocols for safe sharing. Implementing a thorough network security solution may help businesses save money on operating expenses and safeguard themselves from major losses caused by security incidents like data breaches. You may be certain that the system's interfaces, applications, and systems will function correctly if only authorized users are granted access.

Firewalls control the data that enters and leaves a network using predetermined protocols, according to devices that are interacting with one other. Firewalls are essential in everyday computer networking because they streamline traffic and regulate the flow of packets. Firewalls are an essential component of any network security design due to their main role of preventing viruses and application-layer assaults. Network segmentation allows for the identification of many entities inside an organization that provide a comparable function, risk, or status. An enterprise's network is linked to the Internet via a perimeter gateway. Avoiding other networks and organizations that might pose a danger is essential for protecting the systems' resources. By incorporating additional measures such as access control and internal network barriers, organizations might potentially fortify their security configuration.

By controlling which users, groups, and devices may access which parts of a network, access control helps ensure the confidentiality of sensitive data. System administrators may use role-based access control rules and Identity and Access Management systems to ensure that only authorized users and devices are able to access network resources. In every system, remote device access should be considered. Virtual Private

Networks (VPNs) empower users—whether they are hosts, consumers, or employees—to securely access the network from any location or device type. As a matter of course, every host should have a client version of VPN installed or use a web-based client. To further guarantee confidentiality and authenticity, security measures including encryption, endpoint compliance monitoring, and multi-factor authentication are also used.

## III. OBJECTIVE OF RESEARCH

The aim of IDS is to identify different, kinds of malware as early as possible, which cannot be achieved by a traditional firewall. With the increasing volume of computer malware, the development of improved IDSs has become extremely important and has a significant research issues. The objective of this research study on "Exploring the NSL-KDD Dataset: a comprehensive analysis about intrusion detection system", defined as

1- To study the different cased on IDS and current research issues in networking environment.
2- To develop a predictive model and based logistic regression classifier .
3- To evaluate accuracy of predictive model and its efficiency on big data.

## IV. RESEARCH DESIGN AND METHODOLOGY

A statistical model typically used to model a binary dependent variable with the help of logistic function. Another name for the logistic function is a sigmoid function and is given by:

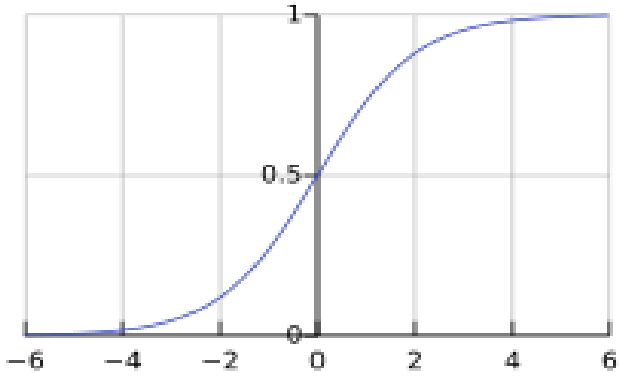


**Fig.1.1:** Sigmoid function

$$F(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{e^{x}+1} \quad \text{.....(1)}$$

This function assists the logistic regression model to squeeze the values from (-k,k) to (0,1). Logistic regression is majorly used for binary classification tasks; however, it can be used for multiclass classification.

The reason behind this is that just like Linear Regression, logistic regression starts from a linear equation. However, this equation consists of log-odds which is further passed through a sigmoid function which squeezes the output of the linear equation to a probability between 0 and 1. And, we can decide a decision boundary and use this probability to conduct classification task. For example, let's assume we are predicting whether it is going to rain tomorrow or not based on the given dataset, and if after applying the logistic model, probability comes out to be 90% then we can surely say that it is highly possible to rain tomorrow. On the other hand, if probability comes out to be 10%, we may say that it is not going to rain tomorrow, and this is how we can transform probabilities to binary.

Since Logistic regression predicts probabilities, we can fit it using likelihood. Therefore, for each training data point x, the predicted class is y. Probability of y is either p if y=1 or 1-p if y=0. Now, the likelihood can be written as:

$$L(\alpha_0, \alpha) = \prod_{I=1}^{n} p(x_i)^{y_i}(1-p(x_i)^{1-y_i} \quad \text{.....(2)}$$

The multiplication can be transformed into a sum by taking the log:

$$l(\alpha_0, \alpha) = \sum_{i=0}^{n} y_i \log p(x_i) + (1 - y_i) log 1 - p(x_i)$$

$$= \sum_{i=0}^{n} \log 1 - p(x_i) + \sum_{i=0}^{n} y_i log \frac{p(x_i)}{1 - p(x_i)} \quad \ldots\ldots(3)$$

Further, after putting the value of p(x):

$$l(\alpha_0, \alpha) = \sum_{i=0}^{n} -\log 1 + e^{\alpha_0 + \alpha} + \sum_{i=0}^{n} y_i(\alpha_0 + \alpha . x_i) \quad \ldots\ldots..(4)$$

The next step is to take a maximum of the above likelihood function because in the case of logistic regression gradient ascent is implemented (opposite of gradient descent).

XG Boost Classifier

It builds decision trees sequentially with each tree attempting to correct the mistakes made by the previous one. The process can be broken down as follows:

1- Start with a base learner: The first model decision tree is trained on the data. In regression tasks this base model simply predict the average of the target variable.
2- Calculate the errors: After training the first tree the errors between the predicted and actual values are calculated.
3- Train the next tree: The next tree is trained on the errors of the previous tree. This step attempts to correct the errors made by the first tree.
4- Repeat the process: This process continues with each new tree trying to correct the errors of the previous trees until a stopping criterion is met.
5- Combine the predictions: The final prediction is the sum of the predictions from all the trees.

## V. RESULTS AND DISCUSSION

5.1 Logistics Predictive Model

| Table1.1: Testing Result of Predictive Model | | | | |
|---|---|---|---|---|
| | precision | recall | f1-score | support |
| 0 | 0.96 | 0.93 | 0.94 | 5828 |
| 1 | 0.94 | 0.97 | 0.95 | 6770 |
| accuracy | | | 0.95 | 12598 |
| macro avg | 0.95 | 0.95 | 0.95 | 12598 |
| weighted avg | 0.95 | 0.95 | 0.95 | 12598 |

The above data analysis report is generated by logistics predictive model and produced the accuracy 0.95% and precision 0.96 and recall 0.93 and F1-score 0.94%. The researcher used the total data 12598 where truly predict ted 6770 where as false predicted 5828(Table1.1).

5.2 Evaluation – XG-Boost Model

| Accuracy | Table 1.2: Evaluation - XGBoost_model | | | |
|---|---|---|---|---|
| | precision | recall | f1-score | support |
| 0 | 1.00 | 1.00 | 1.00 | 5828 |
| 1 | 1.00 | 1.00 | 1.00 | 6770 |
| accuracy | | | 1.00 | 12598 |
| macro avg | 1.00 | 1.00 | 1.00 | 12598 |
| weighted avg | 1.00 | 1.00 | 1.00 | 12598 |

The above data analysis report is generated by : Evaluation – XG- Boost model and produced the accuracy 1% and precision 1% and recall 1 % and F1-score 1%. The researcher used the total data 12598 where truly predict ted 6770 where as false predicted 5828(Table1.2). So we can see that ensemble methods such as xgboost, adaboost, gradient boosts has more accuracy scores over logistic regression in bigger datasets. It doesn't necessary but we will do hyper parameter tuning in order to fit the model with best parameters, i would like to remember that xg-boost has cross-validation has itself

5.3 Hyper-Parameter Tuning: XGB- Classifier

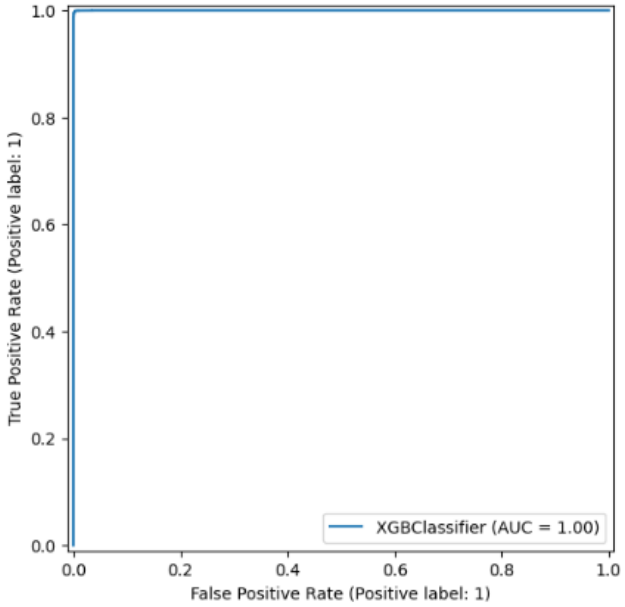


**Fig.1.1:** Hyper-Parameter Tuning: XGB- Classifier

The evaluation datasets play a vital role in the validation of any IDS approach, by allowing us to assess the proposed method's capability in detecting intrusive behavior. The datasets used for network packet analysis in commercial products are not easily available due to privacy issues. However, there are a few publicly available datasets such as DARPA, KDD, NSL-KDD and ADFA-LD and they are widely used as benchmarks. Existing datasets that are used for building and comparative evaluation of IDS, finally the researcher found Hyper-Parameter Tuning: XGB-Classifier accuracy 0.99%.

## VI. SUMMARY AND CONCLUSION

Finally the researcher concluded that the research study "Exploring the NSL-KDD Dataset: A Comprehensive Analysis about Intrusion Detection System (IDS)" are significant and have a great impact of any network suspicious activities. This article presented a review of the research trends in network-based intrusion detection systems (NIDS), their approaches, and the most common datasets used to evaluate IDS Models. The researcher used the supervised machine learning approach logistics and XGB- classifier by using NSL-KDD Dataset. The researcher found that logistic classifier given 0.95% accuracy where as XGBooster Classifier gives the 1.00% accuracy; due to the over fitting the researcher used the hyper parameter tuning XGB classifier and got the 0.99% accuracy. The researcher assured that the developed predictive model is more accurate and efficient to detect the intrusion during the data transmission. The aim of IDS is to identify different, kinds of malware as early as possible, which cannot be achieved by a traditional firewall. With the increasing volume of computer malware, the development of improved IDSs has become extremely important and has a significant research issues.

## REFERENCES

[1]. A. Abbasi, J. Wetzels, W. Bokslag, E. Zambon, and S. Etalle, "On emulation-based network intrusion detection systems," in Research in attacks, intrusions and defenses: 17th international symposium, RAID 2014, Gothenburg, Sweden, September 17–19, 2014. Proceedings, A. Stavrou, H. Bos, and G. Portokalidis, Eds. Cham: Springer International Publishing, 2014, pp. 384–404

[2]. A. A. Aburomman and M. B. Ibne Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system," Appl Soft Comput, vol. 38, pp. 360–372, 2016/01/01/ 2016.

[3]. Adebowale A, Idowu S, Amarachi AA (2013) Comparative study of selected data mining algorithms used for intrusion detection. International Journal of Soft Computing and Engineering (IJSCE) 3(3):237–241.

[4]. Agrawal S, Agrawal J (2015) Survey on anomaly detection using data mining techniques. Procedia Computer Science 60:708–713.

[5]. M. Ahmed, A. Naser Mahmood, and J. Hu, "A survey of network anomaly detection techniques," J Netw Comput Appl, vol. 60, pp. 19–31, 1// 2016.

[6]. A. Alazab, J. Abawajy, M. Hobbs, R. Layton, and A. Khraisat, "Crime toolkits: the Productisation of cybercrime," in 2013 12th IEEE international conference on trust, security and privacy in computing and communications, 2013, pp. 1626–1632.

[7]. A. Alazab, M. Hobbs, J. Abawajy, and M. Alazab, "Using feature selection for intrusion detection system," in 2012 international symposium on communications and information technologies (ISCIT), 2012, pp. 296–301.

[8]. S. Zhao, M. Chandrashekar, Y. Lee, and D. Medhi, —Real-time network anomaly detection system using machine learning,‖ in 2015 11th International Conference on the Design of Reliable Communication Networks (DRCN), Mar. 2015, pp. 267–270. doi: 10.1109/DRCN. 2015.7149025.

[9]. W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, —Multi-level hybrid support vector machine and extreme learning machine based on modified Kmeans for intrusion detection system,‖ Expert Syst. Appl., vol. 67, pp. 296–303, Jan. 2017, doi: 10.1016/j.eswa.2016.09.041.

[10]. Y.-X. Meng, —The practice on using machine learning for network anomaly intrusion detection,‖ in 2011 International Conference on Machine Learning and Cybernetics, Jul. 2011, vol. 2, pp. 576–581. doi: 10.1109/ICMLC.2011.6016798.

[11]. A. Tsiligkaridis and I. Ch. Paschalidis, —Anomaly detection in transportation networks using machine learning techniques,‖ in 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), Nov. 2017, pp. 1–4. doi: 10.1109/URTC.2017.8284194.

[12]. M. E. KarsligEl, A. G. Yavuz, M. A. Güvensan, K. Hanifi, and H. Bank, —Network intrusion detection using machine learning anomaly detection algorithms,‖ in 2017 25th Signal Processing and Communications.

[13]. R. Patgiri, U. Varshney, T. Akutota, and R. Kunde, —An Investigation on Intrusion Detection System Using Machine Learning,‖ in 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Nov. 2018, pp.1684–1691. doi: 10.1109/SSCI.2018.8628676.

[14]. B. S. Bhati, C. S. Rai, B. Balamurugan, and F. Al-Turjman, —An intrusion detection scheme based on the ensemble of discriminant classifiers,‖ Comput. Electr. Eng., vol. 86, p. 106742, Sep. 2020, doi:10.1016/j.compeleceng.2020.106742.

[15]. Shubham Malhotra, Muhammad Saqib, Dipkumar Mehta, and Hassan Tariq. (2023). Efficient Algorithms for Parallel Dynamic Graph Processing: A Study of Techniques and Applications. International Journal of Communication Networks and Information Security (IJCNIS), 15(2), 519–534.

[16]. Divyatmika and M. Sreekesh, —A two-tier network based intrusion detection system architecture using machine learning approach,‖ in 2016 International Conference on Electrical, Electronics, and Optimization Techniques(ICEEOT), Mar. 2016, pp. 42–47. doi: 10.1109/ICEEOT.2016.7755404.

[17]. D. Ashok Kumar and S. R. Venugopalan, —A Novel Algorithm for Network Anomaly Detection Using Adaptive Machine Learning,‖ in Progress in Advanced Computing and Intelligent Engineering, Singapore, 2018, pp. 59–69. doi: 10.1007/978-981-10-6875-1_7.

[18]. A Novel Malware Analysis Framework for Malware Detection and Classification using Machine Learning Approach | Proceedings of the 19th International Conference on Distributed Computing and Networking.‖ https://dl.acm.org/doi/abs/10.1145/3154273.3154326 (accessed Aug. 22).

[19]. T. Kacem, D. Wijesekera, P. Costa, and A. Barreto, —An ADS-B Intrusion Detection System,‖ in 2016 IEEE Trustcom/Big Data

SE/ISPA, Aug. 2016, pp. 544–551. doi: 10.1109/TrustCom.2016.0108.

[20]. M. S. Koli and M. K. Chavan, ―An advanced method for detection of botnet traffic using intrusion detection system,‖ in 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), Mar. 2017, pp. 481–485. doi: 10.1109/ICICCT.2017.7975246.

[21]. P. I. Radoglou-Grammatikis and P. G. Sarigiannidis, ―Flow anomaly based intrusion detection system for Android mobile devices,‖ in 2017 6th International Conference on Modern Circuits and Systems Technologies (MOCAST), May 2017, pp. 1–4. doi: 10.1109/MOCAST.2017.7937625.

[22]. Yerra, S. (2024). The impact of AI-driven data cleansing on supply chain data accuracy and master data management. Smart Computing Systems, 4(1), 187-191. https://doi.org/10.61485/SMCS.27523829/v4n1P1

[23]. K. N. K. Thapa and N. Duraipandian, ―Malicious Traffic classification Using Long Short-Term Memory (LSTM) Model,‖ Wirel. Pers. Commun., vol. 119, no. 3, pp. 2707–2724, Aug. 2021, doi: 10.1007/s11277-021-08359-6.

[24]. Sachin Dixit, & Jagdish Jangid. (2024). Asynchronous SCIM Profile for Security Event Tokens. Journal of Computational Analysis and Applications (JoCAAA), 33(06), 1357–1371. Retrieved from https://eudoxuspress.com/index.php/pub/article/view/1935

[25]. M. Agarwal, S. Purwar, S. Biswas, and S. Nandi, ―Intrusion detection system for PS-Poll DoS attack in 802.11 networks using real time discrete event system,‖ IEEECAA J. Autom. Sin., vol. 4, no. 4, pp. 792–808, 2017, doi:10.1109/JAS.2016.7510178.

[26]. 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), Nov. 2015, pp. 1–6. doi: 10.1109/MilCIS.2015.7348942.

[27]. R. R. Reddy, Y. Ramadevi, and K. V. N. Sunitha, ―Effective discriminant function for intrusion detection using SVM,‖ in 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Sep. 2016, pp. 1148–1153. doi: 10.1109/ICACCI.2016.7732199.

[28]. S. KishorWagh, V. K. Pachghare, and S. R. Kolhe, ―Survey on Intrusion Detection System using Machine Learning Techniques,‖ Int. J. Comput. Appl., vol. 78, no. 16, pp. 30–37, Sep. 2013, doi: 10.5120/13608-1412.

[29]. S. Othman, F. Ba-Alwi, T. Nabeel, and A. Al-Hashida, ―Intrusion detection model using machine learning algorithm on Big Data environment,‖ J. Big Data, vol. 5, Sep. 2018, doi: 10.1186/s40537-018-0145-4.

[30]. S. Layeghy, M. Baktashmotlagh, and M. Portmann, ―DI-NIDS: Domain Invariant Network Intrusion Detection System.‖ arXiv, Oct. 15, 2022. Accessed: Oct. 21, 2022. [Online]. Available: http://arxiv.org/abs/ 2210.08252.