International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. **08** Special Issue **01** | **2023**

Improving Hybrid Classifier Approaches for Anomalies Extraction using IDS Datasets

Dr.VidhyaSathish[1] , Dr.S.Meenakshi[2]

[1]*Assistant Professor, Department of Bachelor of Computer Applications, SDNB Vaishnav College for Women, Chennai-44.*
[2]*Assistant Professor, Department of Bachelor of Computer Applications, SDNB Vaishnav College for Women, Chennai-44.*

ABSTRACT

Objective : To reduce the number of false alarms by improving detection accuracy to the greatest extent possible using supervised classifier approaches and Grey Wolf Search agents. Findings : Existing procedures have an accuracy of detection that ranges from 83 to 99.2% and a false alarm rate of 0.28 to 0.24 percent. They succeeded in lowering false alarm rates, however there is still a discrepancy between increasing high detection accuracy and low false alarm rate. The experimental findings of previous research from the reality of intrusions present in various optimization strategies are discussed in this paper. Methods : The goal of the proposed study was to improve the performance of various supervised machine learning classifier techniques, including Extreme Learning Machine, Multi-Layer Perceptron, and Support Vector Machine, while simultaneously focusing on reducing the false alarm rate from a variety of datasets, including NSLKDD and DARPA99. This was accomplished through experimental analysis using Grey-Wolf Optimization search agents. Additionally, the experiment background had been completed utilizing the MS-Windows operating system's WEKA simulator application. Novelty : The proposed study demonstrated its success in the extraction of anomalies for detection accuracy ranging from 96.97% to 99.95% and false alarm rate ranging from 3.001% to 0.001%, respectively. As a result, the proposed study investigation had demonstrated its success in achieving better results when compared to the obstacles already in place.

Keyword : - Intrusion Detection, Internet of Things, Grey Wolf Optimization, Cyber Security, Machine-Learning techniques.

1. INTRODUCTION

In the area of cyber security, the behaviour of intrusions that use aggressive tactics to use zero-day exploits results in terrible things. Diverse detecting approaches proliferated to overcome problems. Instead of causing harm externally, intrusions harm inside systems with the intention of disseminating malware or obtaining authorized data over a network connection. To put it another way, taking control of hacked technology is the most typical technique to injure someone illegally. In a significant way, viruses, Trojans, and bots have developed to assemble themselves on infected PCs. The problem of over-identification of intrusions arises from the difficulty of achieving the two main goals of a high detection rate and a low false alarm rate. Although the decline in false alarm rates is regrettable, most detection approaches have shown to be effective in reaching these objectives[1][2][3].

Anomalies-based traffic detection in general and Signature-based traffic detection in particular were used to structure detection analyses in order to create full methodologies. The system had been trained to detect known malware using signature-based detection techniques. The importance of these detection techniques encourages researchers to deepen their descriptive knowledge of network traffic analysis. These detection techniques were obstructed, making it impossible to find abnormal network traffic. Anomalies-based detection algorithms, on the other hand, had expanded their study to include finding network anomalies in a range of setups. Neither single-handed nor hybrid approaches were successful in identifying these hazards.

Machine learning classifiers and evolutionary-based strategies were used to produce the prestigious work for both the single-handed neither strategy and hybrid techniques. Two important pieces of information are provided to researchers by an evaluation of significant datasets, which looks at their use and influence on the advancement of intrusion detection systems (IDS) during the past ten years, as well as by a taxonomy of network threats and the tools used to carry out these assaults. The study claims that just 33.3 percent of our threat taxonomy is now covered by IDS research. The accuracy of existing machine learning IDS systems depends on the presence of real-network threats, attack representation, and a significant number of deprecated threats, all of which are conspicuously absent from the datasets at hand [4].

Real-Time Sequential Deep Extreme Learning Machine The RTS-DELM-CSIDS (Cybersecurity Intrusion Detection System) security model [5] was developed to help grade security features according to their importance and to provide a thorough intrusion detection framework centred on the key traits. 70 percent of the NSL-KDD data were used for training (103,962 samples), 30 percent for validation (103,962 samples), and 44,554 samples were randomly selected. In order to eliminate data discrepancies and safeguard data from mistakes, data was processed beforehand. The RTS-DELM-CSIDS scans numerous hidden layers, including hidden neurons, and activation processes for harmful activity or infiltration. Additionally, to accurately predict the effectiveness of the system, the experiment counted a certain number of neurons in buried levels of a network and used a range of active activities. The RTS-DELM-CSIDS framework has an accuracy of 96.22 percent and a missing rate of 3.27 percent, which is better than other methods such the self-organizing map (SOM), which has an accuracy of 75.5 percent, ANN-based IDS, which has an accuracy of 81.2 percent, and generative adversarial networks (GANs), which has an accuracy of 86.5 percent.

The Grasshopper Optimization Algorithm[6] (GOA) is used to enhance and more precisely learn ANNs in order to decrease intrusion detection error rate. The GOAMLP method selects advantageous parameters like weight and bias in order to lower the intrusion detection error of the neural network. The implementation in the MATLAB programme and use of the KDD and UNSW datasets demonstrate the excellent accuracy with which the suggested method detects abnormal, hostile traffic and attacks. The GOAMLP method outperforms and outperforms current state-of-the-art techniques for network intrusion detection, including RF, XGBoost, and embedded learning of Artificial Neural Networks (ANN) with Butterfly Optimization Algorithm (BOA), Harris Hawks Optimization (HHO), and Black Window Optimization (BWO) algorithms.

In comparison to embedded learning approaches based on KDD, which have a detection accuracy of 95.41 percent, and UNSW datasets, which have a detection accuracy of 98.88 percent, the GOA methodology has superior accuracy, sensitivity, and specificity. With the aim of improving intrusion detection accuracy and detection rate while decreasing processing time in the WSN environment by reducing false alarm rates and the amount of features generated by IDSs, the modified Grey Wolf Optimizer and SVM [7] for enhanced intrusion detection system were developed. Three wolves, five wolves, and seven wolves were employed by the GWOSVM-IDS to determine the right number of wolves. The outcomes of this experiment are demonstrated using the NSL-KDD99. With 7 wolves and a time period of 69.6 hours, GWOSVM-IDS has a detection accuracy of 96 percent and a false alarm rate of 0.03 percent. On the other hand, the GWOSVM-IDS 5 and GWOSVM-IDS 3 wolves had detection accuracy of 92% and 79%, respectively, with false alarm rates of 0.096% and 0.24%. For 12 and 27 features, GWOSVM-IDS with 5 wolves and GWOSVM-IDS with 3 wolves took 74.4 and 86.4 hours, respectively.

Due to the effectiveness of big data applications, many machine learning applications are being converted to deep learning models. The work on enhancing the performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Datasets [8] revealed the success of deep learning algorithms in spotting tiny sample assaults. Six different machine learning classifier models were applied to an up-to-date datasets to enhance the performance of the Machine Learning Intrusion Detection System. The datasets CSE-CIC-IDS2018 was used to construct Decision Tree, Random Forest, K-Nearest Neighbour, Adaboost, Gradient Boosting, and Linear Discriminant Analysis models. With respect to accuracy, the figures were 99.66%, 99.21%, 98.52%, 99.69%, 99.11%, and 90.80%, respectively, while the error rate was 0.34%, 0.79%, 1.14%, 0.31%, 0.89%, and 9.20%.

1.1 System model

The existing techniques discussed here [1-8] analyse IDS based on Machine Learning algorithms and evolutionary-based methodologies, which have sparked interest in building notable IDS in recent years. Based on this existing fact, the goal of this proposed experimentation is to create hybrid classification techniques that combine the additive combined approach of the Grey Wolf Optimization algorithm with the Support Vector Machine, Extreme Learning Machine, and Multi Layer Perceptron classifier techniques to improve classifier detection in a short amount of time. During preprocess and attribute feature selection, the proposed hybrid classifier techniques each being well trained to show the betterment classification results using 'confusion metrics' evaluation method as shown in Figure.1.

In addition, each classifier technique (Multi Layer Perceptron, Support Vector Machine, Extreme Learning Machine, Logistic, Sequential Minimal Optimization with Radial Basis Kernel Function (SMORBF), and PolyKernel Function) was trained separately with two different datasets (NSLKDD & DARPA99 1-14 week). The requirements achieved by each classifier technique were validated using a random set of datasets instances. The number of iterations with each classifier technique was examined to validate the performance during the experiment. Tables 1 and 2 show the results of each classifier technique's analysis of two different datasets based on this.

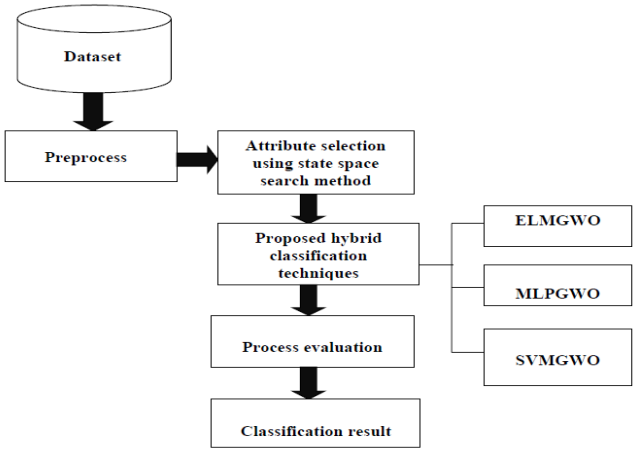

Figure 1. Architecture for Enhanced Hybrid Classification for Intrusion Detection

Investigating intrusion cases utilizing modified NSLKDD and DARPA99 1-14 week datasets is the stated purpose of the effort. The outcomes of the modified KDDCUP99 intrusion datasets were then contrasted with these. In order to show the proposed hybrid classifier's high detection accuracy and low false positive rate, performance of the proposed hybrid classifier techniques will be compared and evaluated among one another, as well as against the results of the modified KDDCUP99 intrusion datasets and existing techniques.

DARPA99[9] is one of the most well-known datasets used in the IDS sector. Actually, the DARPA 99 datasets is really a better version of the DARPA 98 datasets. A testbed was developed to generate both legitimate and malicious traffic. An estimated five million records make up the DARPA99 Datasets. Twenty-three attributes are listed in each record. In this work, a case study from the flow-based Intrusion Detection Evaluation (IDEVAL) datasets was selected to examine the Secure Shell Handshake (SSH protocol). An open source protocol called SSH makes it possible to log into a machine from a distance. Additionally, tunnelling advances the arbitrary TCP ports across the secured channel, such as Skype, which is indicated for file transfers, between the user and the remote system.To identify the encrypted and non-encrypted traffic traces that will be gathered, the following techniques are applied to this datasets: ELMGWO, MLPGWO, SVMGWO, ELM, SVM (Polykernel, Radial Basis Function), and MLP. Table 1 lists the DARPA99 datasets characteristics that were tested using ELMGWO, MLPGWO, SVMGWO, ELM, SVM (Polykernel, Radial Basis Function), and MLP.

Table 1. Classifier analysis using Darpa99 1-14week datasets

Classifier	No.Of. Instances (Time Taken To Build Model)	Evaluation Metric Used[Full Trainiing Set]	No.Of. Attributes	No.Of. Iterations	Correctly Classified Instances	Incorrectly Classified Instances
ELM	38064 (11.49 sec)	%5 F.T.S	23	1st	36160 (99.9972%)	1 (0.0028%)
ELM	38064 (10.52 sec)	% 10 F.T.S	23	2nd	34257 (99.9971%)	1 (0.0029%)
ELM	38064 (4.12 sec)	5fold Cross Validation F.T.S	23	3rd	38063 (99.9974%)	1 (0.0026%)
ELMGW	38064 (11.46 sec)	% 5 F.T.S	23	1st	34257 (99.9971%)	1 (0.0029%)
ELMGW	38064 (3.76 sec)	5fold C.V. F.T.S	23	2nd	34257 (99.9971%)	1 (0.0029%)
ELMGW	38064 (3.45 sec)	10fold C.V. F.T.S	23	3rd	34257 (99.9971%)	1 (0.0029%)
LOGISTIC	38064 (4.78 sec)	%5 F.T.S	23	1st	36160 (99.9972%)	1 (0.0028%)
LOGISTIC	38064 (4.38 sec)	5fold C.V. F.T.S	23	2nd	38063 (99.9974%)	1 (0.0026%)
MLP	38064 (839.67 sec)	%5 F.T.S	23	1st	36160 (99.9972%)	1 (0.0028%)
MLP	38064 (821.54 sec)	5fold C.V. F.T.S	23	2nd	38063 (99.9974%)	1 (0.0026%)
MLPGW	38064 (789.56 sec)	%5 F.T.S	23	1st	36160 (99.9972%)	1 (0.0028%)
MLPGW	38064 (902.47 sec)	5fold C.V. F.T.S	23	2nd	38063 (99.9974%)	1 (0.0026%)
MLPGW	38064 (4.28 sec)	%50 F.T.S	23	3rd	19031 (99.9947%)	1 (0.0053%)
MLPGW	38064 (7.41 sec)	%25 F.T.S	23	4th	28547 (99.9965%)	1 (0.0035%)
SMO POLYKERNEL	38064 (1.4 sec)	%5 F.T.S	23	1st	36160 (99.9972%)	1 (0.0028%)
SMO POLYKERNEL	38064 (0.53 sec)	5fold C.V. F.T.S	23	2nd	38063 (99.9974%)	1 (0.0026%)
SMO	38064	%27 F.T.S	23	3rd	27786	1

POLYKERNEL	(0.94 sec)				(99.9964%)	(0.0036%)
SMO RBF KERNEL	38064 (2.01 sec)	%27 F.T.S	23	4th	27786 (99.9964%)	1 (0.0036%)
SVMGW POLYKERNEL	38064 (1.11 sec)	%5 F.T.S	23	1st	36160 (99.9972%)	1 (0.0028%)
SVMGW POLYKERNEL	38064 (0.64 sec)	5fold C.V. F.T.S	23	2nd	38063 (99.9974%)	1 (0.0026%)
SVMGW NORMALIZED KERNEL	38064 (24.06 sec)	%47 F.T.S	23	3rd	20173 (99.995%)	1 (0.005%)

Table 1 lists the evaluation metric (whole training set), the number of occurrences, and the number of examples used to build the model. The DARPA99 (1–14) week datasets, attributes, and correctly and incorrectly identified cases were studied using a classifier. Around 38064 instances were chosen at random to be tested. Cross Fold validation and Percentage Split presence of Full Training Set are the two types of metrics used to complete the evaluation in order to confirm that the results are the same when the evaluation is performed in two different metric modes. The datasets is iterated two or three times for each algorithmic model, using each of the 23 attributes. The experiment's findings show that the suggested study is significantly more accurate at classifying abnormalities than ELM, SVM, MLP, and LOGISTIC (ELMGW achieved in the range of 99.9971 percent, MLPGW achieved in the range of 99.9965 percent, and SVMGW achieved in the range of 99.9995 percent).

The benchmark datasets for intrusion detection was the NSL-KDD datasets[7] [10]. Because there are no duplicates in the train set, the classifiers in the NSL-KDD datasets do not yield biassed outputs. Since there are no duplicate records in the suggested test sets, learners' performance won't be hampered and detection rates will be higher. The small number of records chosen at each level of difficulty is inversely correlated with the percentage of records in the KDD datasets. The following methods are used on this datasets to distinguish between the encrypted and non-encrypted traffic traces that will be gathered: ELMGWO, MLPGWO, SVMGWO (Polykernel, Radial Basis Function), ELM, and MLP. The NSL-KDD datasets features used in the experimental versions of the ELMGWO, MLPGWO, SVMGWO (Polykernel, Radial Basis Function), ELM, and MLP are listed in Table 2.

Table 2. Classifier analysis using NSL-KDD dataset

Classifier	No.Of. Instances (Time Taken To Build Model)	Evaluation Metric Used[Full Trainiing Set]	No.Of. Attributes	No.Of. Iterations	Correctly Classified Instances	Incorrectly Classified Instances
ELM	25192 (3.23 sec)	%10 F.T.S	42	1st	20792 (91.7038%)	1881 (8.2962%)
ELM	25192 (3.07 sec)	10 fold C.V. F.T.S	42	2nd	23013 (91.3504%)	2179 (8.6496%)
ELM	25192 (22.82 sec)	5 fold C.V. F.T.S	42	3rd	24546 (97.4357%)	646 (2.7662%)
ELMGW	25192 (22.47 sec)	%5 F.T.S	42	1st	23270 (97.2338%	662 (2.7662%)

					)	
ELMGW	25192 (21.89 sec)	5 fold C.V. F.T.S	42	2nd	24546 (97.4357%)	646 (2.7662%)
ELMGW	25192 (3.04 sec)	%10 F.T.S	42	3rd	20792 (91.7038%)	1881 (8.2962%)
ELMGW	25192 (2.84 sec)	10 fold C.V. F.T.S	42	4th	23013 (91.3504%)	2179 (8.6496%)
LOGISTIC	25192 (3.5sec)	%10 F.T.S	42	1st	20792 (91.7038%)	1881 (8.2962%)
LOGISTIC	25192 (3.12 sec)	10 fold C.V. F.T.S	42	2nd	23013 (91.3504%)	2179 (8.6496%)
MLP	25192 (99.78 sec)	%10 F.T.S	42	1st	22075 (97.3625%)	598 (2.6375%)
MLP	25192 (47.56 sec)	10 fold C.V. F.T.S	42	2nd	13182 (52.3261%)	12010 (47.6739%)
MLPGW	25192 (34.95 sec)	%10 F.T.S	42	1st	21849 (96.3657%)	824 (3.6343%)
MLPGW	25192 (25.39 sec)	10 fold C.V. F.T.S	42	2nd	24436 (96.999%)	756 (3.001%)
SVMGW POLYKERNEL	25192 (1036.78 sec)	%5 F.T.S	42	1st	23172 (96.8243%)	760 (3.1757%)
SVMGW POLYKERNEL	25192 (1406.26 sec)	%10 F.T.S	42	2nd	21987 (96.9744%)	686 (3.0256%)

Table 2 lists the experimental results, including the number of examples used to build the model, the evaluation measure (complete training set), and the number of instances. Examining the classifiers, features, and iterations of the NSL-KDD datasets as well as examples that were successfully and randomly categorized 25192 instances in total were randomly selected for testing. In order to confirm that the output is the same when the evaluation is performed in two independent measure modes, the evaluation is completed with two types of metrics, namely Cross Fold validation and Percentage Split presence of Full Training Set. For each algorithmic model, the datasets is iterated two or three times, using all 42 attributes each time.

In the accurate classification of instances, the experimentation results of the proposed MLPGW reached in the range of 96.99 percent, ELMGW attained in the range of 97.43 percent, and SVMGW attained in the range of 96.97 percent. The NSL-KDD dataset achieved detection accuracy of 99.89 percent with a false alarm rate of 1-2 percent, and the UNSW-NB15 dataset achieved 91.86 percent detection accuracy, according to the proposed experimentation results of both the DARPA99 and NSL-KDD datasets compared with the hybrid machine learning method[10]. As indicated, the proposed testing produced a detection rate that was noticeably better than earlier results, with a false alarm rate of about 0.002%.

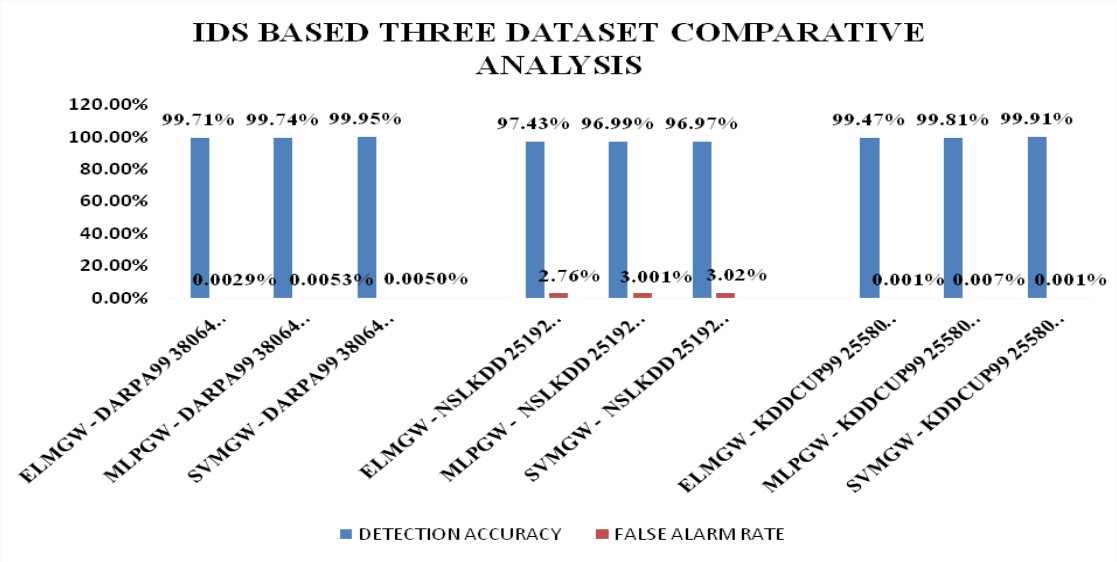

Figure 2. represents IDS Based three Benchmark Datasets Analysis

The modified set of instances around 38064 improved the detection accuracy of ELMGW, MLPGW, and SVMGW from DARPA99 to 99.71 percent, 99.47 percent, 99.74 percent, 99.23 percent, and 99.95 percent, respectively. The false alarm rate was 0.0029 percent, 0.026 percent, 0.0053 percent, 0.0077 percent, and 0.0005 percent. In a separate mode of experimentation, using NSL-KDD with a modified set of instances near 25192, detection accuracy was found to range between 91.35 and 97.43 percent, 96.36 to 96.99 percent, and 96.82 to 96.97 percent, with false alarm rates of 8.29 to 2.76 percent, 3.63 to 3.001 percent, and 3.17 to 3.02 percent, respectively. The proposed results were compared to the KDDCUP99 modified dataset values with the range of 25580 instances[11] for ELMGW, MLPGW, and SVMGW at 98.96% - 99.47%, 99.28% - 99.81%, 99.908% - 99.83% as detection accuracy and their respective false positives at the range of 0.007%, 0.001%, and 0.001% as shown in Figure 2. to achieve the betterment results in achieving.

CONCLUSION

The experimental investigation of anomalies extraction using ELMGW, MLPGW, and SVMGW from the deployment of NSLKDD; DARPA99 datasets finds that enhanced evident of anomalies extraction traces as compared to the current obstacle. With a false alarm rate of 0.0029, 0.0053, and 0.0050%, respectively, the proposed anomaly extraction approaches have detection accuracy of 99.71 percent for ELMGW, 99.74 percent for MLPGW, and 99.95 percent for SVMGW. The same optimization techniques will need to be used on other datasets in the future as an upgrade.

REFERENCES

[1] Sarika Choudharya,_, Nishtha Kesswanib. Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 Datasets using Deep Learning in IoT,International Conference on Computational Intelligence and Data Science (ICCIDS 2019), Procedia Computer Science 167:1561-1573, DOI:10.1016/j.procs.2020.03.367.

[2] Safaldin, M., Otair,M., and Abualigah, L. Improved Binary Grey Wolf Optimizer and SVM for Intrusion Detection System in Wireless Sensor Networks. Journal of Ambient Intelligence and Humanized Computing, Springer-Verlag, Heidelberg, 2020, https://doi.org/10.1007/s12652-020-02228-z.

[3] Abdullah Alzaqebah, Ibrahim Aljarah , Omar Al-Kadi, and Robertas Damaševi˘cius. A Modifified Grey Wolf Optimization Algorithm for an Intrusion Detection System, *Mathematics* 2022, *10*, 999. https://doi.org/10.3390/math10060999.

[4] Hanan Hindy, David Brosset , Ethan Bayne , Amar Seeam, Christos Tachtatzis, Robert Atkinson And Xavier Bellekens,A Taxonomy of Network Threats and the Effect of Current Datasets on Intrusion Detection Systems, IEEE ACCESS, VOLUME 8, 2020, 104650-104675. **Electronic ISSN:** 2169-3536, **DOI:** 10.1109/ACCESS.2020.3000179.

[5] Amir Haider, Muhammad Adnan Khan, Abdur Rehman, Muhib Ur Rahman and Hyung Seok Kim. A Real-Time Sequential Deep Extreme Learning Machine Cybersecurity Intrusion Detection System,Computers, Materials & Continua, Vol.66, No.2, 2021, *pp.1785-1798*, doi:10.32604/cmc.2020.013910.

[6] Shadi Moghanian , Farshid Bagheri Saravi ,Giti Javidi, And Ehsan o. Sheybani GOAMLP: Network Intrusion Detection With Multilayer Perceptron and Grasshopper Optimization Algorithm, 2020, 215202-215213, IEEE Access 8:215202 - 215213, DOI:10.1109/ACCESS.2020.3040740.

[7] Safaldin, M., Otair,M., and Abualigah, L., 2020. Improved Binary Grey Wolf Optimizer and SVM for Intrusion Detection System in Wireless Sensor Networks. Journal of Ambient Intelligence and Humanized Computing, Springer-Verlag, Heidelberg, https://doi.org/10.1007/s12652-020-02228-z.

[8] Karatas, G., Demir, O., and Sahingoz, O.K., 2019. Increasing the performance of Machine Learning - Based IDS on an Imbalanced and up-to-date Dataset. IEEE Access, 4, pp.1-13. DOI:10.1109/ACCESS.2020.2973219

[9] Ansam Khraisat , Iqbal Gondal, Peter Vamplew and Joarder Kamruzzaman, Survey of intrusion detection systems: techniques, datasets and challenges, Cybersecurity (2019) 2:20 https://doi.org/10.1186/s42400-019-0038-7

[10] Achmad Akbar Megantara and Tohari Ahmad,A hybrid machine learning method for increasing the performance of network intrusion detection systems,J Big Data (2021) 8:142 https://doi.org/10.1186/s40537-021-00531-w.

[11] Vidhya S, Khader, P.S.A., 2018. Enhanced Hybrid Classifier Techniques Using Grey Wolf Optimization for Improved Detection Accuracy. http://hdl.handle.net/10603/218662