Shahid Chamran University

Of Ahvaz

"MSc" Thesis

Anomaly based intrusion detection system using Golden Eagle optimizer and improved random forest model

By:

Riyam Ibrahim Abdulwahid

Supervisor:

Dr. Mohammad Javad Rashti

September 2022

Acknowledgements

Abstract

With the increasing security threat and amount of network throughput, the study of intrusion detection systems (IDSs) has received many attentions throughout the computer science domain. Network data inspection manual classification is a task which is repetitive, takes time, expensive. Mechanism of IDS is too effective for finding anomalies and attacks of network. Techniques of IDS based on Anomaly are the worthy technology for protecting the systems of target as well as networks against malicious activities. IDS controls packets of network for detecting malicious activities. Normally such packets have many features that most of them are not repetitive and relevant also they are able to curtail IDS performance. So, this is essential for using techniques of feature selection for choosing the optimal features' subset. Here, combination of Golden Eagle Optimizer (GEO) and Grey Wolf Optimization (GWO) is used to extract relevant IoT network features. The extracted features are related to the improved random forest (IRF) classifier to achieve high attack detection accuracy. The proposed method is evaluated using the benchmark dataset, namely, Network Security Laboratory-Knowledge Discovery in Databases (NSL-KDD). Studies show that the accuracy for the proposed two-class method were 97.23%, which has improved 9.9% compared to the other methods.

Keywords: intrusion detection systems, Grey Wolf Optimization (GWO), Golden Eagle Optimizer (GEO), feature selection, feature extraction.

Contents

Abstract **3**

CHAPTER ONE INTRODUCTION **8**

1-1- Introduction 9

1-2- Problem Definition 10

1-3- Challenges 12

1-4- Thesis Objectives 12

1-5- Thesis outline 12

CHAPTER TWO DEFINITIONS OF BACSIC CONCEPTS. **13**

2-1- Intrusion detection system (IDS) 14

2-1-1- Based on location 14

2-1-2- Based on detection methods 14

2-2- Categories of intrusion detection systems 15

2-2-1- Signature Based Detection 15

2-2-2- Anomaly Based Detection 16

2-3- Anomaly detection 17

2-4- IDS technologies types 18

2-5- Anomaly detection techniques 20

2-6- Related work about anomaly-based intrusion detection systems 23

CHAPTER THREE PROPOSED METHOD. **29**

3-1- Methodology 29

3-2- collection module of Data 30

3-3- preprocessing of Data 30

3-4- selection module of Feature 30

3-4-1- Grey Wolf Optimization (GWO) 30

3-4-2- Golden Eagle Optimizer (GEO) 32

3-5- Training module 33
3-5-1- Random forest classifier 33
3-6- Knowledgebase module 34
3-7- Testing module 34
3-8- Alert module 34

CHAPTER FOUR RESULT EVALUATION. 35
4-1- Dataset 36
4-1-1- Attack classification 37
4-2- Parameters' Initialization 38
4-3- Evaluation criteria 39
4-4- The results' evaluation 41

CHAPTER FIVE CONCLUSION. 44
5-1- Conclusion 45
5-2- Future works 46

CHAPTER SIX REFERENCES 47
CHAPTER SEVEN APPENDIX 51

List of Figures

Figure 2- 1- anomaly detection [11] 17

Figure 2- 2- A simple example of anomalies in a 2-dimensional data set [12]..... 18

Figure 3- 1- Proposed framework for intrusion detection 29

Figure 4- 1- The proposed model in terms of average accuracy 42

List of Tables

Table 4- 1- The classification of types of attacks 37
Table 4- 2- Initial values for the parameters in the proposed method 38
Table 4- 3- The confusion table [29] 39
Table 4-4-The comparison results of the proposed method and the base paper ….42

Chapter 1

Introduction

1-1- Introduction

Detecting bad intrusions of network for many years has been a subject of study. Since scientists of data are able to appreciate, but, while the issue scale increases with the magnitude order, sometimes present approaches are not efficient anymore; issue is various sufficiently that needs novel solution. Since network traffic amount has increased via magnitude orders, intrusion detection domain has caused in reinvent across techniques of large data. The IDS controls either networks / the other systems for anomalous/ bad manners. Fulling preventative technologies like privilege of user, strong authentication, firewalls, IDSs have become the important enterprise IT security management section [1].

IDS approaches are categorized as anomaly detection-based IDS, misuse detection-based IDS, and hybrid approaches. IDS frameworks based on signature depend on familiar security threats signatures, however generating whole attacks and bad manners database of signature are able to be the time-consuming and arduous activity. In addition, IDS approaches based on signature are not able to cope with novel attacks where the traffic is encrypted and signatures are not familiar. In other words, IDS models based on anomaly monitor based on usual users' manner profiles and are able to detect unleashed attacks newly. But, this is hard for determining also keeping whole normal manners in wide organizations. In order to take advantage of both groups when dealing with the lacks, multiple models of IDS attempt in efficiently combining methods of detection of anomaly and misuse [2].

Methods of Machine learning for detection of intrusion have been researched by the investigators for across 20 years. Wide network telemetry amount and the other security data sorts has made issue of intrusion detection able for the methods of machine learning. A lot of novel detection systems of commercial intrusion apply algorithms-based machine learning as the detection strategy section (such as Cisco

Stealthwatch and Microsoft Azure Sentinel of security platforms). In general, such methods fall under intrusion detection technique anomaly detection group [3].

1-2- Problem Definition

Over the past decades the network security has changed with threats becoming far more complex moving from basic attacks against one device to network intrusion attacks against organizations networks. A network intrusion attack is defined as any use of a computer network that compromises network security. Intruders try to gain unauthorized access to files or privileges, modify and destroy the data, or render the computer network unreliable. The goal of intrusion detection is to build a system which would scan network activities and generate alerts if either a specific attack occurred or an anomaly in the network behavior detected [4].

Intrusion detection as a topic and field of research, can generally be divided into two different methods, signature detection and anomaly detection. Signature detection are methods where already known attacks are identified by information stored in the system, so called "signatures". Anomaly detection are the methods that define what "normal" traffic is, and then classify everything that falls out of that category as "anomalies" and therefore are potential intrusions. Unsurprisingly, signature detection is very effective against the known attacks in its data base, but is unable to detect any other intrusions. Anomaly detection on the other hand, might detect new types of intrusions if trained correctly, but runs a higher risk of creating false positives. Anomaly detection also tends to have a lower true positive rate than signature detection of known attacks. It is a generally accepted option that a good intrusion detection system should rely on both signature detection and anomaly detection [5].

IDS based on Anomaly recognizes the malicious tasks with generating profile for arranged tasks and comparing to usual profile. IDS requires monitoring as well as investigating packets of network. The packets have too many attributes which define

them like the type of protocol, destination or source address of IP address and so on. A lot of attributes are extra and non-related that make data classification/ analysis hard and degrade IDS performance. Therefore, it comes methods of feature selection (FS) importance. FS is the processing stage of data that targets in eliminating whole feasible extra and non-related attributes from underlying vector of feature/dataset also decreasing requirement of time as well as storage for processing data and raising performance of system. FS is taken as the full combinatorial issue of optimization NP. Creating whole feasible possible features subsets as well as assessing them is not possible for huge set of data. Currently, techniques of Meta-heuristic have become well-known to solve various issues particularly to solve the issues of feature selection because of their ability in obtaining the optimum or near-optimal solution in the reasonable time [6].

Highly supported features are achieved with Hybrid Golden Eagle Optimizer - Grey Wolf Optimization (GWO– GEO) for the efficient detection of intrusion. the first time improved for IoT intrusion detection of network is Hybrid GWO–GEO that pursues the important function of fitness for eliminating extra and non-related features. Oversampling has been applied for addressing imbalanced data problem.

Modern sets of data include vibrant data which is gathered from a lot of sensors and devices of IoT, making them as the data with high-dimension. The data with high-dimension involves several non-related features, ruing performance of model. module of feature selection assists in choosing related subset of feature (manually /automatically) for improving accuracy, developing performance reducing time of training, preventing/reducing overfitting, decreasing size of data.

Grey Wolf Optimization (GWO) [7] and Golden Eagle Optimizer (GEO) integration [8] is applied for extracting the related features of IoT network. extracted features are fed to the developed classifier of random forest (IRF) for obtaining the high accuracy of attack detection.

1-3- Challenges

The high level of false alarms that are generated reduces IDS performance against cyber-attacks and specially, makes problem for the tasks of a security analyst, results in intrusion management process to be more expensive in terms of computing.

1-4- Thesis Objectives

The main objective of this study is the intrusion detection using a bagging classifier such as Random Forest optimized by GWO– GEO. The contributions are as follows: contributions include:

1. For the efficient detection of intrusion, highly supported features are achieved with Hybrid GWO–GEO. first time improved for intrusion detection of IoT network is Hybrid GWO– GEO that pursues the important function of fitness for eliminating extra and non-related features.
2. Oversampling has been utilized for addressing imbalanced data problem.

1-5- Thesis outline

First, in chapter 2 we will describe the basic concepts, and we review the related work and in the end. In chapter 3 we describe proposed approach. In chapter 4, the experimental results have been purpose and finally we make the conclusion in chapter 5.

Chapter 2

Definitions of basic concepts

2-1- Intrusion detection system (IDS)

Intrusion detection systems are the Mechanism for tracking networking environments intrusion. That describes bad calculating network resources usage. That is the basic element to detect the attacks based on Internet which might grouped in network and host based [9].

2-1-1- Based on location

- Network Intrusion Detection System (IDS)

NIDs analyze information flow among calculating areas like traffic of network. Such IDS are observing the traffic on specific network points. That impacts at power point of network such as router, gateway via traffic of network.

- Host based Intrusion Detection System (HIDS)

HIDS are installed on personal devices/ system in network. IDS analyses outgoing and incoming packet of data from real system. server attempts in recognizing attacks with snooping transactions, files of registration, traffic and so on. This is greater rather than the IDS of Network as the comparison in order to detect bad works for the real system. The IDS sort impacts personal system/ host by unwanted shifts of configuration are detected [9].

2-1-2- Based on detection methods

Detection of Signature– the method of detection is according to matching of pattern. This compares packets of data by familiar bad order. The id is simple for developing when you will find network works' kind in order to be recognized. Such IDSs which might have too high accuracy while identifying low false positives and familiar attacks. The IDSs' kinds are not able to figure out new attacks. Popular supervised classifier like Naïve Bayes, artificial neural network based on back propagation, Decision Dree are utilized for matching/Detection of Signature.

Detection of Anomaly–the method of detection is according to the approach of Geometric also utilized for detecting the unobserved attacks. Unsupervised algorithm like artificial neural network is utilized in detection of Anomaly. The approach is the centralized way which acts on baseline network manner concept. The manner of baseline network that is learned, specified also described with administrators of network system. IDS based on Anomaly is able to detect the unfamiliar attacks when the system is not updated [9].

2-2- Categories of intrusion detection systems

2-2-1- Signature Based Detection

Detection of Signature includes traffic of searching network for the malicious bytes or packet order. This method Basic benefic is that the signatures are too simple for developing and identifying when we know what manner of network we are attempting in order to recognize. For example, we may apply the signature which searches the specific strings in exploit specific vulnerability of buffer overflow. Happenings created by IDS-based signature are able to communicate alert cause. Since the matching of pattern is able to be more efficiently performed on new systems, therefore power amount required for performing the matching is minimal for the set of rule. For instance, when protected system is just communicate through SMTP, DNS, ICMP, whole of the other signatures are able to be forgotten. Such signature engines' restrictions are that they just detect the attacks which signatures are saved in base of data before; the signature should be generated for each attack; also new attacks are not able in order to be detected. The method is able to be deceived simply due to that they are just according to the orderly matching of string as well as expressions. Such mechanisms just search the strings in forwarding packets across the wire. In addition, signatures well perform against just stabled the pattern of manner, they fail for coping with the attacks which are generated by worm/ human by manner features which are self-correcting. Detection based on Signature

does not well perform while a user applies the improved technologies such as encrypted channels of data, generators of nop, payload encoders. Systems of efficiency based on signature are reduced broadly since this must generate the modern signature for each variation. Since signatures continue to raise, performance of system engine reduces. Because of that, a lot of engines of intrusion detection are employed on systems by hybrid processors as well as hybrid cards of Gigabit network. Developers of IDS improve novel signatures before an attacker does, therefore avoiding new attacks on system. Novel signatures difference creation speed among attackers and developers assign system efficiency [10].

2-2-2- Anomaly Based Detection

Detection based on anomaly is according to describing manner of network. Manner of network is related to the manner which is described before, after that this is adapted that this begins the happening in detection of anomaly. Adapted manner of network is learned/ prepared with network administrators specifications.

Basic step in describing manner of network is capability of IDS engine for cutting via different protocols at whole levels. Engine should be able in processing protocols as well as understanding the aim. However, the analysis of protocol is expensive in computation order, advantages creates as it raising set of rule assists in less alarms of false positive. Basic anomaly detection disadvantage is describing the set of rule. system efficiency relies on how well this is tested and performed on whole protocols. Process of Rule defining is influenced by different protocols applied by different vendors. On the other hand, typical protocols cause the defining of rule as a hard work. detailed knowledge on adapted manner of network require in order to be improved by administrators for detection in order to happen accurately. However, while rules are described and the protocol is generated the systems of anomaly detection well performs. While bad user manner falls under adapted manner, this goes unnoticed. The work like traversal of directory on the server of targeted

vulnerable that complies by protocol of network, simply goes unnoticed as that does not begin each restriction flags of bandwidth, out-of-protocol, payload. Main detection based on anomaly benefit across the engines based on signature is that the modern attack for what the signature does not exist is not able to be detected while this falls out of usual patterns of traffic. It is seen while systems detect novel automated worms. While modern system is infected by the worm, typically that triggers to scan the other systems of vulnerable at the rate of accelerated completing network by bad traffic, therefore resulting in TCP connection event/ rule of bandwidth abnormality [10].

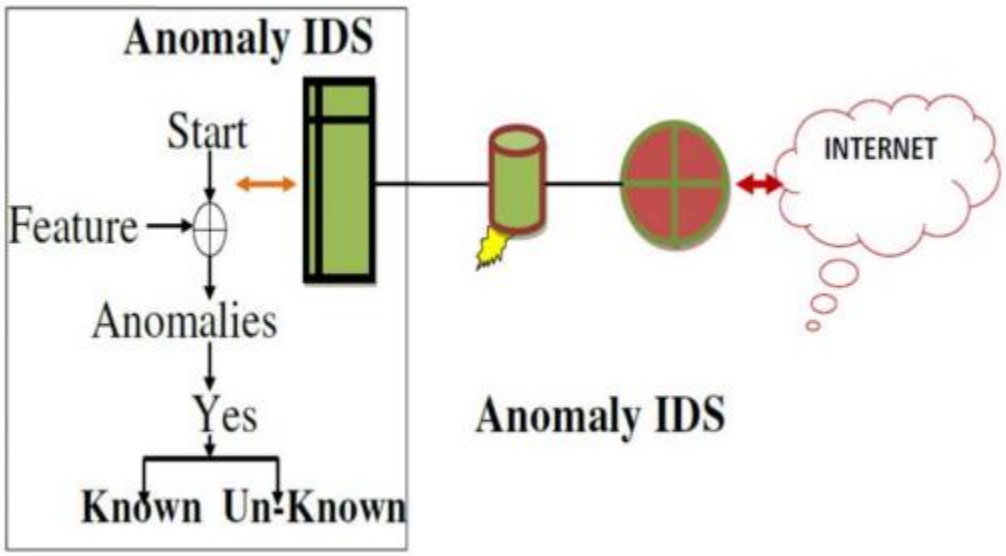

Figure 2- 1- anomaly detection [11]

Since the IDS kind is created encompassing costumers' profiles, this is in addition called as identification based on profile in order Figure. 2-1.

2-3- Anomaly detection

Detection of Anomaly is the method which is applied for detecting uncommon patterns which do not adapt the typical manner. Detection of Anomaly has a lot of apps in different fields differs from detection of intrusion to the health of system controlling and from detection of fraud in transactions of credit card to detect the fault in areas of operating.

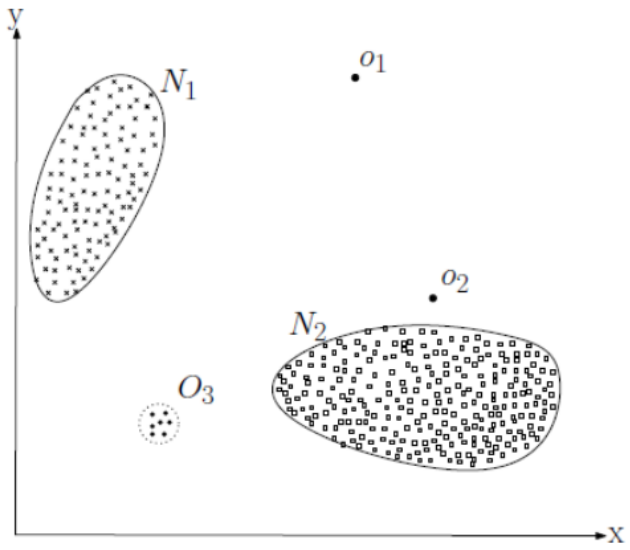

Figure 2- 2- A simple example of anomalies in a 2-dimensional data set [12].

Fig. 2-2 shows the simple anomalies instance in 2D set of data. data has 2 usual areas which are grouped, N2 and N1. The Points which are far from such areas sufficiently are anomalies. Bad works' traces applied for reflecting in data, careful data analysis shows intruder existence. Data anomaly traces have the usual features that makes detection of anomaly feasible to the analyst.

2-4- IDS technologies types

Todays, a lot of IDS technology's sorts exist. We group technologies in 4 levels based on the place they are employed to the works of inspect suspicious, also what kinds of happening they are able to identify. The four levels are as: Wireless-based IDS (WIDS), Host-based IDS (HIDS), Mixed IDS (MIDS), Network based IDS (NIDS), Network Behavior Analysis (NBA). Host-based IDS gathers and controls features for hosts including the suspicious works, running public services of servers, sensitive information. NIDS gets the traffic of network at particular parts of network via sensors, also in order that analyzes apps as well as protocols' activities for

recognizing suspicious incidents. WIDS is like the NIDS, however that gets the wireless traffic of network like wireless mesh networks, ad hoc networks, wireless sensor networks. In addition, system of NBA inspects the traffic of network for identifying attacks by unexpected flows of traffic. Accepting the hybrid technologies like MIDS is able to meet an aim for the more correct and complete detection [13]. IDS elements are the factor and sensor that former is normally utilized for systems of NBA, NIDS, WIDS for controlling networks, HIDS applies latter for monitoring as well as analyzing the tasks. Two factor and sensor are able to transmit data to Database Server (DS) and Management Server (MS) that MS is the centralized device to process the captured incidents, DS is only the repository saving information of happening. Furthermore, two network architectures sorts exist. One of them refers to Managed Network (MN), the isolated network which is employed for software management of security for showing information of IDS from intruders. MN raises additional costs of hardware also brings the real inconveniences for administrators. The other one refers to Standard Network (SN) that is the general network with no protection. A path in improving security of Standard Network is creating the virtual isolated network with configuring the virtual local network of environment. In other words, most of the technologies of IDS present 4 typical abilities to maintain security such as logging, prevention, collecting of information as well as detection. Collection of Information gathers the information about networks or hosts from the works which are seen. Logging, associated data of logging for detected happenings are able to be applied for validating checked events as well as alerts. in most IDSs, Methods of Detection normally requires significant tuning for getting the higher accuracy [13].

2-5- Anomaly detection techniques

Techniques of Anomaly detection are applied widely in issues of intrusion detection. Such techniques are able to identify two familiar as well as unfamiliar attacks that makes them more sufficient rather than the techniques based on signature. Additionally, such techniques are effective for creating novel signatures in IDSs based on signature. A lot of techniques of anomaly detection exist that are recognized in literature. All of them are applied in creating usual network profile also finding the unusual patterns.

statistical models-based Intrusion detection techniques believe that the anomalies are samples of data which weakly fit in the statistical model. They create the parametric traffic model and use the statistical test for grouping data of traffic. The model is created from statistical historical network traffic data parameters. Outputs of statistical test measure the similarity. When the latter is wider rather than the threshold which is described before, data of traffic are taken into consideration as bad ones. There are two kinds of Statistical models. At First, we distinguish univariate techniques of change detection that analyze 1 parameter each time like median, mean, standard deviation, etc. Here, one is able to place methods of control chart such as algorithm of CUSUM, average of geometric moving, Shewhart, exponential weighted moving average (EWMA). control chart of CUSUM takes whole historical network data values into consideration for detecting intrusions in present traffic which is used on data of network traffic. It is performed with computing cumulative sum, showed by Ct, from the deviations of recorded values mean in window of training. For detecting intrusions, control chart of CUSUM applies statistic Ct which accumulates for mostly high observed process distance amounts to (μ0 + k) that k is the presented threshold. Some criteria are taken into consideration as compare whole the methods. Such methods' performance is measured in false positives rate (FPR)terms, needed time for detecting an attack,

small and progressive intrusions detection ability. one of the basic techniques applied in statistical process control (SPC) is Algorithms of Control charts [14]. Between such algorithms, one is able to place Shewhart utilized in monitoring average of process. EWMA in the parameters of process is the typically applied chart of control for deviation detection. control chart of EWMA has been considered in a lot of papers of research, especially in field of SPC also it is applied broadly in networks for detection of anomalies and intrusion. Cisar et al used algorithm of EWMA for intrusion detection of network. An algorithm checks whether characteristics of traffic exceed the threshold which is defined before in considered time like packets number. Sklavounos et al assessed EWMA algorithm performance to detect various cyber network intrusions kinds like Probe, U2R, distributed denial of service (DDoS), R2L. Sklavounos et al modeled the model of intrusion detection for detecting TCP packets intrusion of R2L. this model applies 2 charts of control: CUSUM, EWMA. The tests were implemented on dataset of NSL-KDD, two charts provided great outcomes in accuracy of detection terms.

Writers modeled the IDS for detecting attacks of DoS by applying chart of CUSUM. Both techniques were tested and provided: first one just utilized source bytes of UDP in statistical test, however second technique checks sources bytes of ICMP and UDP in CUSUM parameters computation. second group considers multivariate algorithms of intrusion detection that are used to more complicated shifts like non-additive multidimensional data shifts. Such methods investigate links among two/more criteria. They contain multivariate methods like models of forecasting as well as time-series [like model of autoregressive integrated moving average (ARIMA)], ratio of likelihood, principle component analysis (PCA). Soule et al presented the technique for Kalman filter-based traffic anomalies detection. Some literature studies proved IDSs time series models ability. Fouladi et al created the series of time from received flows and packets of traffic. After that, 4 statistical time series measures are utilized and computed for detecting attacks of DDoS. Tests

illustrated that measure of skewness performed better rather than the other parameters to detect the attacks of DDoS.

Writers modeled the model of ARIMA for detecting the attack of DDoS. first, packets and source addresses of IP amount every minute are extracted from received traffic of network for building series of time. Next, writers used model of ARIMA for predicting upcoming packets amount. At last, network traffic classification is done by using 2 rules: chaotic behavior repeatability, packets number ratio growth in source IP addresses amount. Main ARIMA model restriction is that they use the stationary series of time. Upcoming traffic amounts are considered in order to be the linear past amounts' function. So, nonlinear patterns are not able in order to be gotten by the model. Additionally, because of dynamic users' manners' shift, sometimes this is hard to assign whether last modeled design will be efficient in future. Statistical models have some benefits in real world. first, they result the quantitative measure showing a degree to which assessed sample of data is anomalous. Additionally, they do not need the last attacks of network/ usual knowledge of system manner as they are able to learn expected pattern of network from received data of traffic. At last, they are able to develop rates of precision as well as accuracy of detection, also decrease false alarms applying suitable thresholds. In spite of the benefits of them, statistical methods have several restrictions. Adjusting statistical test parameter amounts is the hard work. always a trade-off exists among the alarms of FN and FP. Most of the statistical designs consider that controlled data is the process of quasi-stationary/ stationary. Situation is not satisfied always in real world. Additionally, sometimes deviation from expected manner is reported the long time after deviation start. As a result, sometimes this is difficult for parametrically modeling data, particularly in multidimensional data [14].

2-6- Related work about anomaly-based intrusion detection systems

In [15], an enhanced anomaly-based IDS model based on multi-objective grey wolf optimization (GWO) algorithm was proposed. The GWO algorithm was employed as a feature selection mechanism to identify the most relevant features from the dataset that contribute to high classification accuracy. Furthermore, support vector machine was used to estimate the capability of selected features in predicting the attacks accurately. Moreover, 20% of NSL–KDD dataset was used to demonstrate effectiveness of the proposed approach through different attack scenarios. The experimental result revealed that the proposed approach obtains classification accuracy of (93.64%, 91.01%, 57.72%, 53.7%) for DoS, Probe, R2L, and U2R attack respectively. Finally, the proposed approach was compared with other existing approaches and achieves significant result.

In [16], the detection framework of network intrusion as well as the scheme have been presented for area of IoT network. Scheme was performed in programming language of python. Schemed was assessed datasets of CICIDS-2017, KDDCup99, NSL–KDD by utilizing recall, accuracy, F1-score, Precision. Set of data is accessible in 8 files of CSV integrated in starting the training the balanced model of GWO–PSO–RF IDS for network of IoT. balanced model of GWO–PSO–RF NIDS was assessed on 25 percent of various sets of data. results were compared to algorithms of logistic regression (LR), Naive Bayesian (NB), decision tree (DT) by multiple GWO–PSO. This has been illustrated that the balanced training of data eliminates the low prediction issue of class when the unbalanced training of set of data overlooked low classes in prediction. Presented scheme has been compared to the current techniques also obtained the highest average 99.66 percent accuracy for whole the taken sets of data. Therefore, presented balanced scheme of GWO–PSO–RF IDS for network of IoT addressing biasing issue to more frequently happening records as well as raises DR for the low attacks' levels.

In [17], the writers present novel approach with integrating methods of Adaptive Grasshopper Optimization Algorithm (AGOA) as well as Ensemble of Feature Selection (EFS) named EFSAGOA that are able to assist in recognizing attacks sorts. Basically the method of EFS is used for ranking feature to select high ranked features' subset in proposed technique. After that, AGOA is developed for assigning significant features from decreased sets of data which is able to help predicting traffic treat of networks. Moreover, GOA adaptive treat utilizes in deciding whether record shows the anomaly or not, varying from several techniques acquainted in the literature. AGOA utilizes Support Vector Machine (SVM) as the function of fitness for selecting widely effective attributes as well as maximizing performance of classification. Additionally, this is used for optimizing SVM classifier factor of penalty (C), tube size (ϵ), kernel parameter (σ). EFSAGOA performance has been assessed on the new data of intrusion like ISCX 2012. results of experiments illustrate that presented technique outperforms also achieves low rate of false alarm, high rate of detection, accuracy in comparison with the other modern methods in data of ISCX 2012.

In [18], ensemble feature selection (EFS) and grasshopper optimization algorithm (GOA) combination named EFSGOA is improved. At first, method of EFS is used for ranking attributes to choose top related features subset. Then, GOA is applied for recognizing important attributes from achieved decreased set of features which is created with method of EFS which is able to help in determining attack sort. Moreover, GOA uses the SVM as the function of fitness for obtaining noteworthy features also optimizing the factor of penalty, SVM tube size parameters, parameter of kernel to maximize performance of classification. Results of the test show that technique of EFSGOA has outperformed also achieved the high rate of detection 99.69 percent as well as accuracy 99.98 percent and low false rate of alarm 0.07 in NSL-KDD and high rate of detection 99.26 percent, accuracy of 99.89 percent as well as low false rate of alarm of 0.097 in data of KDD Cup 99. In addition, presented

technique has succeeded to obtain the higher performance in comparison with the other modern methods in case of detection rate, CPU time, accuracy, rate of false alarm.

In [19], baselines of showcases multiclass classification by utilizing the various algorithms of neural networks and ML to distinguish the traffic of legitimate network from obfuscated and direct intrusions of network. The paper obtains baselines from the dataset of Tunneling Obfuscations and Advanced Security Network Metrics. Set of data which is captured obfuscated and legitimate malicious communications of TCP on the chosen vulnerable services of network. Hybrid NIDS of classification can distinguish direct and obfuscated intrusion of network by up to 95 percent accuracy.

In [20], the writers have used preprocessing in KDD 99 and gathered the set of data by utilizing the gain of information. The writers called gathered set of data as NUM15 since several extra data and features are in addition the point that reduces time of processing as well as IDS of performance. Then, Snort and naive Bayes are utilized for classifying results of compression as well as training machine in parallel scheme. Multiple scheme integrates detection of signature and anomaly which is able to obtain network anomaly detection. results illustrate that present multiple scheme is able to raise accuracy as well as determining the new intrusions.

In [21], the writers check efforts of network intrusion by ML schemes based on anomaly for presenting the better protection rather than conventional models based on misuse. Two schemes named as convolutional neural network and ensemble learning were created and performed on the set of data which is collected from environment of institutional production and real-world. For showing reliability and validity of schemes, they were used to benchmarking data set of UNSW-NB15. Attack sort was restricted in probing attacks for keeping study scope controllable. Results show that high rates of accuracy, model of CNN slowly being more suitable.

In [22], firstly autoencoders (AEs) are applied for decreasing original data dimension also the multiple model integrating GWO and PSO is presented for optimizing parameters of SVM. The technique integrates 2 algorithms of optimization also chooses optimum values of parameter based on locally increased particles for training classifier. Here, UNSW-NB15 dataset as well as benchmark dataset of NSL-KDD are utilized for assessing presented scheme also the scheme is separately compared to the other techniques of classification. Results of test illustrate that the multiple model of optimization has better detection accuracy performance also presents the great rate of detection as well as false rate of alarm.

In [23], stream mining of data is combined by the IDS for doing the specified task. task is distinguishing significant, successfully covered up information in less time amount. Test concentrates on developing IDS effectiveness by utilizing presented Stacked Autoencoder Hoeffding Tree approach (SAE-HT) applying the Darwinian Particle Swarm Optimization (DPSO) for selection of feature. test is implemented in dataset of NSL_KDD and significant attributes are achieved by utilizing DPSO and classification is done utilizing presented method of SAE-HT. presented method obtains the higher accuracy of 97.7 percent while compared to all other modern methods. This is seen that presented method raises rate of detection and accuracy then decreasing false rate of alarm.

In [24], presents the novel reliable multiple technique for the anomaly network-based IDS (A-NIDS) by utilizing algorithms of AdaBoost and artificial bee colony (ABC) for gaining the high detection rate (DR) by low false positive rate (FPR). Algorithm of ABC is utilized for the selection of feature as well as AdaBoost are utilized for assessing and classifying features. simulation Results on datasets of ISCXIDS2012 and NSL-KDD confirm that the reliable multiple technique has outstanding difference from the other IDS that are obtained based on similar set of data. This has variously showed better performance in various scenarios based on

attacks. The technique rate if detection and accuracy has been developed compared to in legendary techniques.

Chapter 3

Proposed Method

3-1- Methodology

The effective framework of intrusion detection has been presented for detecting the intrusions effectively in area of network. proposed framework technique by sub-modules of it has been illustrated in Figure 3-1. seven modules exist which are utilized in proposed framework technique based on below:

- module collection of Data
- module preprocessing of Data
- module selection of Feature
- module of Training
- module of Knowledgebase
- module of Testing
- module of Alert

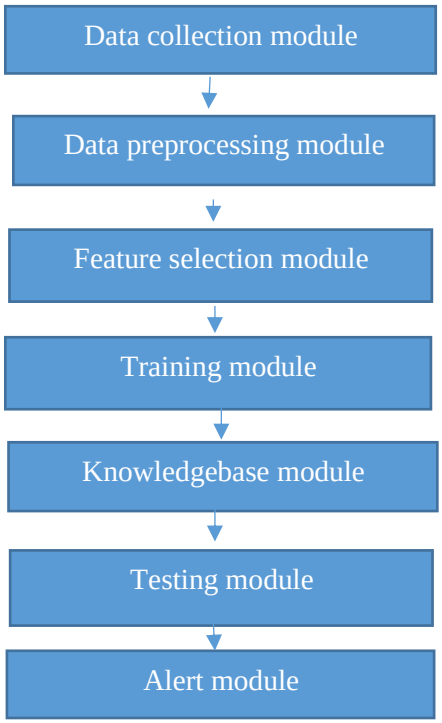

Figure 3- 1- Proposed framework for intrusion detection

3-2- collection module of Data

Here, set of data is able to be prepared from flow of network by utilizing various tools of logging and the model is able to be trained for identifying an intrusion. Accessible datasets publicly related to Network like CICIDS-2017, KDDCup99, NSL–KDD are able to be taken into consideration for collection of data.

3-3- preprocessing module of Data

Data of real world are in various shapes like videos, audio, structured, images, unstructured. Here, by applying numerilization data are easily normalized, cleaned, transformed, encoded, balanced, and normalization to be parsed via ML process machine.

3-4- selection module of Feature

selection module of feature assists in choosing the related subset of feature (manually/ automatically) for preventing/reducing overfitting, improving performance, reduces training time, develops accuracy, decreasing size of data. preprocessed data have been used to selection module of feature. GWO and GEO integration as hybrid GWO– GEO has been used for giving the better selection of feature subset.

3-4-1- Grey Wolf Optimization (GWO)

Grey wolf (Canis lupus) belongs to Canidae family. Grey wolves are considered as apex predators, meaning that they are at the top of the food chain. Grey wolves mostly prefer to live in a pack. The group size is 5–12 on average. Of particular interest is that they have a very strict social dominant hierarchy. In addition to the social hierarchy of wolves, group hunting is another interesting social behavior of

grey wolves. Mathematically for modeling wolves' social hierarchy while forming GWO, we take the fittest solution as alpha (a). accordingly, third and second solutions are known as respectively delta (d) and beta (b), candidate solutions rest are estimated in order to be omega (x). hunting (optimization) is guided by a, b, and d in algorithm of GWO. Wolves of x pursue such three wolves.

- **Encircling prey**

Grey wolves encircle prey in hunt as mentioned earlier. Mathematically, for modeling encircling treat the equations below are presented:

$$\vec{D} = |\vec{C} \cdot \vec{X}_p(t) - \vec{X}(t)|$$

$$\vec{X}(t+1) = \vec{X}_p(t) - \vec{A} \cdot \vec{D}$$

That t shows present iteration, $\vec{A}$, $\vec{C}$ are vectors of coefficient, $\vec{X}$p is prey position vector, $\vec{X}$ shows grey wolf position vector. vectors $\vec{A}$ and $\vec{C}$ are computed as below:

$$\vec{A} = 2\vec{a} \cdot \vec{r}_1 - \vec{a}$$

$$\vec{C} = 2 \cdot \vec{r}_2$$

That ~a elements are reduced linearly from 2 -0 through iterations course and r1, r2 are random vectors in [0, 1].

- **Hunting**

Wolves of Grey are able in identifying prey location also encircle them. Usually, hunt is guided by alpha. Delta and beta may participate occasionally in hunting. but, in the search space of abstract we have not any opinion on optimum location (prey). Mathematically for simulating grey wolves hunting treat, we assume that alpha (best candidate solution) beta, delta have better information on potential prey location. Hence, we store first three of the best solutions achieved till now and oblige other

agents of search (such as omegas) for updating the locations based on the best search agents' location. in this case formulas below are presented.

$$\vec{D}_\alpha = |\vec{C}_1 \cdot \vec{X}_\alpha - \vec{X}|, \vec{D}_\beta = |\vec{C}_2 \cdot \vec{X}_\beta - \vec{X}|, \vec{D}_\delta = |\vec{C}_3 \cdot \vec{X}_\delta - \vec{X}|$$

$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1 \cdot (\vec{D}_\alpha), \vec{X}_2 = \vec{X}_\beta - \vec{A}_2 \cdot (\vec{D}_\beta), \vec{X}_3 = \vec{X}_\delta - \vec{A}_3 \cdot (\vec{D}_\delta)$$

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3}$$

3-4-2- Golden Eagle Optimizer (GEO)

GWO disadvantages like slow convergence, poor searching ability of local minima, low accuracy is dealing by utilizing best performance indicator value of Gbk via GEO in GWO update Eq position. Hence, it utilized the hybrid algorithm of GWO–GEO for extracting the best possible features.

GEO is according to golden eagles spiral motion. Every golden eagle memorizes the best position that this has so far visited. eagle has the attraction to attack prey also the cruise for searching for better food at the same time. In every iteration, every golden eagle chooses another golden eagle prey randomly also circles around f the best location visited so far by golden eagle. golden eagle is able to select circling f memory of it; so, we have. $f \in \{1,2,\ldots,PopSize\}$.

- **Selection of Prey**

every golden eagle should select the prey for performing operations of attack and cruise in each iteration. prey is formed as the best solution recognized with golden eagles' flock in GEO. Every golden eagle can memorize the best solution that this has recognized till now. Every agent of search chooses the prey of aim from all flock memory in every iteration. vectors of cruise and Attack for every golden eagle are computed based on chosen prey. memory is updated If novel location (calculated via attack and cruise vectors) is better rather than last memory position. Here, every

memory prey is mapped/ determined to one and only one golden eagle. every golden eagle implements operations of cruise and attack on chosen prey.

- **Attack (exploitation)**

attack is able to be shaped through the beginning from present golden eagle position as well as ending in prey location in memory of eagle. golden eagle Vector of attack is able to be computed through equation of i.

$$\vec{A}_i = \vec{X}_f^* - \vec{X}_i$$

Where $\vec{A}$i is the attack vector of eagle i , $\vec{X}$f, is the best location (prey) visited so far by eagle f, and $\vec{X}$i is the current position of eagle i. Since the attack vector guides the population of golden eagles toward the best- i visited locations, it highlights the exploitation phase in GEO.

Taken function of objective is every agent values squared sum (feature) in loop. Function of target is presented as following:

$$F_1(x) = \sum_{i=0}^{n} x_i^2.$$

3-5- Training module

Selected features data which are balanced by utilizing method of data-balancing in this module. balanced data work as the input to chosen classifier. classifier selected might be the classifier based on ML.

3-5-1- Random forest classifier

Classifier of Bagging is one kind of method of ensemble which is called as aggregation of bootstrap. Hybrid models of base (M1, M2, … Mn) are integrated.

Every scheme is presented various records instances by utilizing row sampling by the replacement. Several records might repeat in instances presented to models in row sampling with replacement. Classifier of voting is utilized to aggregate models outputs to achieve the decision. One sort of classifier of bagging that a lot of trees of decision are utilized as hybrid models is Random forest. Every tree of decision is according to input given the column and row sampling. Tree of decision has an issue which this has high variance and low bias. It means tree has the better performance on step of training but weak performance on step of testing. Model of voting decreases variance from high-low as decision does not just rely on the specific tree, this relies on hybrid trees' voting.

3-6- Knowledgebase module

It includes the rules' set for induction such as statements of "if–then". This updated with model and module of training and novel attacks are able to be found. This makes the decisions by applying the rules/ knowledge/ reality.

3-7- Testing module

Module of testing applies data of testing on a scheme for assessing performance of scheme to detect the IoT network environment intrusion according to different parameters like recall or detection rate (DR), false alarm rate (FAR), F1-score, accuracy, Precision.

3-8- Alert module

Alert is sent to administrator of network in recognizing attack/ intrusion. alert includes IP of attacker and victim, port of destination and source, detection result. alerted intrusions are saved in central log.

Chapter 4

Result Evaluation

4-1- Dataset

knowledge discovery and data mining (KDD) of 1999 The CUP is a standard dataset used by researchers for performing simulations on intrusion detection systems (IDS) so that focus more on the technical core and objectivity of IDS performance measurement. The network is made with packets of transmission control protocol (TCP) which begin and end at the specific time among streams of data from one source address of Internet Protocol (IP) to the other goal address of target IP under the specific protocol. Every network is described as attack or normal to have the specific attack kind. Data utilized in here is dataset of NSL-KDD that is the developed KDD CUP 99 DARPA version improved by Lincoln laboratories of MIT. Laboratories of Lincoln simulate area to obtain 9 raw TCP data weeks for local-area network (LAN) also mimic the common air force network of US. Additionally, they performed LAN in a way that this was for instance real space of air force also simulates several attacks. However, RPA and KDD (University of California) DA datasets integration absorbs attention, the ability for reflecting real-world situations widely raised question. Therefore, in this study to have a meaningful research, we will utilize dataset of NSL-KDD which is prepared by Information Security Centre of Excellence (ISCX) at Computer Science School, University of New Brunswick, Canada. data in dataset of NSL-KDD is labeled as one of 24 various attacks' kind or normal. Additionally, 24 attacks are shared in 4 categories: Remote to Local control (R2L), User to Root (U2R), denial of service (DoS), probe [28]. Attacks' kinds based on dataset are briefed in Table 4.1.

In this study, 10% KDD CUP 1999 dataset is used having 494,021 records, 41 attributes (features) and one last attribute for the label. 41 attributes used include 9 basic types, 13 types of content and the rest is the traffic type.

4-1-1- Attack classification

In the data processing step, it is divided into 5 classes of 23 types of attacks in the 10% KDD CUP 1999 dataset. The same as information shown in the table below:

Table 4- 1- The classification of types of attacks

Class	Types of attacks
Normal	Normal records
DoS	Back, land, Neptune, pod, smurf, teardrop
PROBE	Ipsweep, nmap, portsweep, satan
U2R	buffer_overflow, loadmodule, perl, rootkit
R2L	ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster

- **Denial-of-service (DoS**): it is an attack that causes the server to receive the busy processing requests, so it cannot provide the services required by legitimate users. For example, when a user wants to request a Hypertext Transfer Protocol (HTML) service from a server through a web browser, it is not possible to access the service due to the denial-of-service. Back, land, Neptune, pod, smurf, teardrop are types of attacks that fall into the DoS group and such attacks emphasize the period or duration of the attack on the host computer network.

- **PROBE** is an attack in which the attacker scans the network to find and know the security hole. An attacker who already knows the vulnerabilities in the server connected to network can abuse the server. Social engineering techniques are widely used in this type of attack.
- **User-to-Root (U2R),** a legitimate user misuses the system on the server to use it as root or administrator. Buffer overflow is a conventional method used in User-to-Root attacks.
- **Remote-to-Local (R2L)** is an attack performed by the attacker through network on a server for searching security flaws and obtaining a user account for login.

4-2- Parameters' Initialization

In order to evaluate the quality of proposed algorithm, we adjusted the parameters according to the parameters of the existing papers. In this way, 100 were considered for Max_iteration and 20 were considered for SearchAgents_no. In Table 4-2, it shows the settings for parameters of the proposed method.

Table 4- 2- Initial values for the parameters in the proposed method

Parameters	Values
Max_iteration	100
SearchAgents_no	20
Problem dimension	Problem dimension Number of features in the data
Search domain	[0 1]
AttackPropensity	[0.5 , 2]
CruisePropensity	[1 , 0.5]

4-3- Evaluation criteria

In order to evaluate the performance of proposed method model, a comparative analysis with the base paper has been performed using several performance criteria. MATLAB 2020b was used to implement the simulations. The used performance metrics are accuracy, recall, precision and F1 score.

In this study, the performance will be examined. Performance is calculated using accuracy, recall, and f-measure, which are used to check for true and false samples. Such a measurement can be shown using the probability table in Table 4.3.

Table 4- 3- The confusion table [29]

		The predicted tag	
	Predicted / Actual	The legal connection	Intrusion
The actual tag	The legal connection	*True Negative (TN)*	*False Positive (FP)*
	Intrusion	*False Negative (FN)*	*True Positive (TP)*

- True Negative (TN): The number of legal connections that is correctly predicted.
- False Positive (FP): The number of legal connections that is incorrectly predicted.
- False negative (FN): The number of intrusion connections that is not correctly predicted.
- True Positive (TP): The number of intrusion connections that is correctly predicted.

Accuracy (ACC): The rate of samples that are correctly predicted.

$$ACC = \frac{TN + TP}{FP + FN + TP + TN}$$

(4-1)

Recall refers to the test's ability to correctly detect attacks on the network (connections that are actually intrusion). The recall of test method is the ratio of number of connections correctly predicted to the total number of connections that are actually intrusion. Mathematically, this can be expressed as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100$$

(4-2)

Precision checks the difference between the measured value and the actual value. Good accuracy is not the reason for precision, but precision is impossible to be good without accuracy. To have high precision, it is possible to obtain a sufficient number of counts from a measurement with good accuracy and use the appropriate statistical method.

$$\text{Precision} = \frac{TP}{TP+\text{FP}} \times 100$$

(4-3)

The F1 score is a machine learning metric that can be used in classification models. Precision and Recall are the two building blocks of the F1 score. The goal of the F1 score is to combine the precision and recall metrics into a single metric.

$$\text{F1 score} = 2 * \frac{Precision*Recall}{Precision+Recall} \times 100$$

(4-4)

Where FP, TN are true positive and negative numbers, respectively. A complete intrusion detection method must be 100% accurate, while having 0% false positive rate (FPR), which indicates that it can detect all possible attacks without error (incorrect classification), which is very difficult and probably impossible in real environments.

4-4- The results' evaluation

In order to evaluate the proposed method, accuracy is calculated for each sample. In order to provide an overall accuracy in different models, an average accuracy can be considered, which is computed by summing the accuracy of each sample and dividing the results by the total number of samples.

The experiment carried out for the proposed approach uses the NSLKDD dataset, and the results obtained were satisfying. The following configurations are used for performing our analysis:

- In Hardware: 12 GB RAM, 1.80 GHz (8 CPU), Intel core i7 and intel motherboard.
- In Software: 64-bit windows 10 and Matlab 2020b.
- Data set: NSLKDD dataset.

The application of GWO–GEO along with the improved Random Forest worked well in comparison with existing techniques like SVM [25], Naïve Bayes[26], and Decision tree[27]. The tabular form is presented below for the Performance, Accuracy rate (%), and precision rate (%), and recall rate (%), and F1 score rate (%) for different approaches:

Table 4- 4- The comparison results of the proposed method and the base paper

Metric	Accuracy	Precision	Recall	F1 score
SVM [25]	84.73%	-	-	-
Naïve Bayes[26]	87.33%	76.83%	85.52%	-
Decision Tree[27]	83.24%	-	-	-
GWO–GEO with improved Random Forest	97.23%	98.50%	98.88%	98.69%

The table given above gives a numerical representation of the obtained values from the experiment. The precision found in our proposed approach is very high as of 98.50%. As well, the accuracy obtained is much higher than previous algorithms. Also, the recall taken for the performance is high than other algorithms.

The average accuracy is computed after calculating the existing models discussed in section 4.4 and the results are shown in Figure 4.1.

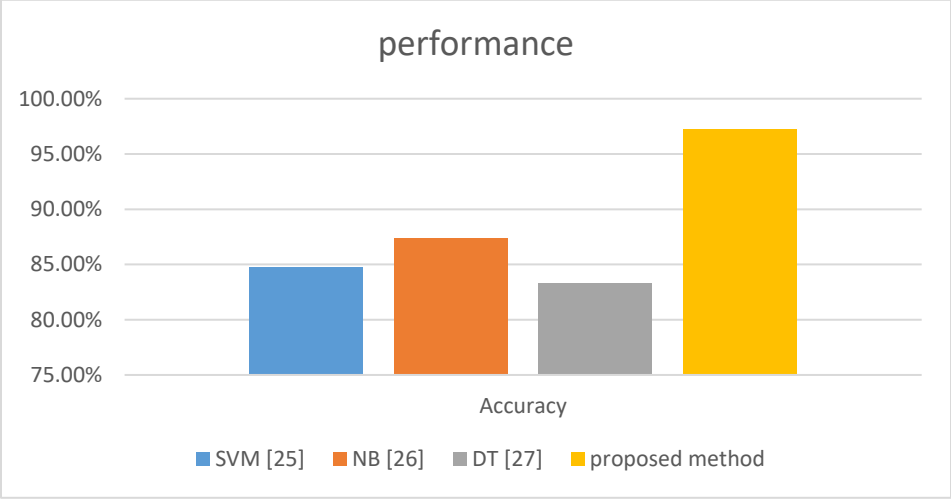

Figure 4- 1- The proposed model in terms of average accuracy

As can be seen, this proposed model has reported an accuracy of 97.23%. The classification efficiency of the proposed method was compared with the existing methods. Table 4.4 provides a comparison for accuracy, recall, precision, and F1 score for two-class classification for the proposed method and the other methods.

A similar dataset is used for all the algorithms presented in Table 4.4. The dataset used is NSLKDD, which is described in section 4.1.

In order to classify two-class classification, tuples are described as normal and attack classes. The accuracy for the proposed two-class method were 97.23%, which has improved 9.9% compared to the other methods

Chapter 5

Conclusion

5-1- Conclusion

Network and data security are some of the most important things for an agency at this time. Various types of attacks that occur through the internet against networks and data encourage agencies to implement various systems to detect and prevent attacks that occur. One system that is often used to detect attacks is intrusion detection system (IDS). IDS is a system used to automate the process of detecting suspicious activity in the network and analyze the possibility of attacks in these activities. There are several methods used in IDS to detect, including anomaly detection and misuse detection. Anomaly detection is detection by comparing the state of an existing activity with the state when a normal activity, while misuse detection is detection by matching the activity pattern with a pattern contained in a database that has been previously defined. Apart from these two methods, several studies have been carried out to conduct detection, prediction, or classification using data mining algorithms. In this study, Grey Wolf Optimization (GWO) and Golden Eagle Optimizer (GEO) integration is applied for extracting the related features of IoT network. extracted features are fed to the developed classifier of random forest (IRF) for obtaining the high accuracy of attack detection. The proposed method is evaluated using the benchmark dataset, namely, Network Security Laboratory-Knowledge Discovery in Databases (NSL-KDD). Studies show that the accuracy for the proposed two-class method were 97.23%, which has improved 9.9% compared to the other methods.

5-2- Future works

The proposed method, like all supervised learners, needs to label the training items. Future work is needed to limit the training rate, which is required to deploy machine learning models for intrusion detection systems. One of the applications of this method is power systems, the topology of power systems can vary significantly in the network. However, the physical behavior of power systems is definite and can be simulated accurately. Some methods are needed to help supervised learning by predicting behaviors based on the physical simulation of the power system. Remote protection is an example of power transmission monitoring and control technology that requires special adjustment for each sample. Engineers perform short circuit simulations in order to obtain remote protection of relay settings. Similar simulation approaches are required to assist supervised learning for IDS power systems.

Future researches are able to be determined for adapting as well as developing improved framework for being effective for specified IoT as well as security problems based on smart city. The other work of future would be controlling datasets based on larger-scale IDS for presenting the comprehensive system of analysis.

6

References

[1]. Imra Salo, Fadi, Mohammadnoor Injadat, Ali Bou Nassif, Abdallah Shami, and Aleksander Essex. "Data mining techniques in intrusion detection systems: A systematic literature review." IEEE Access 6 (2018): 56046-56058.

[2]. Mohammadi, Mokhtar, Tarik A. Rashid, Sarkhel H. Taher Karim, Adil Hussain Mohammed Aldalwie, Quan Thanh Tho, Moazam Bidaki, Amir Masoud Rahmani, and Mehdi Hosseinzadeh. "A comprehensive survey and taxonomy of the SVM-based intrusion detection systems." Journal of Network and Computer Applications 178 (2021): 102983.

[3]. Gamage, Sunanda, and Jagath Samarabandu. "Deep learning methods in network intrusion detection: A survey and an objective comparison." Journal of Network and Computer Applications 169 (2020): 102767.

[4]. Protić, Danijela. "Anomaly-based intrusion detection: Feature selection and normalization influence to the machine learning models accuracy." European Journal of Formal Sciences and Engineering 3, no. 1 (2020): 1-9.

[5]. Wester, Philip. "Anomaly-based intrusion detection using Tree Augmented Naive Bayes Classifier." (2021).

[6]. Mahboob, Amir Soltany, and Mohammad Reza Ostadi Moghaddam. "An Anomaly-based Intrusion Detection System Using Butterfly Optimization Algorithm." In 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), pp. 1-6. IEEE, 2020.

[7]. Mirjalili, Seyedali, Seyed Mohammad Mirjalili, and Andrew Lewis. "Grey wolf optimizer." Advances in engineering software 69 (2014): 46-61.

[8]. Mohammadi-Balani, Abdolkarim, Mahmoud Dehghan Nayeri, Adel Azar, and Mohammadreza Taghizadeh-Yazdi. "Golden eagle optimizer: A nature-inspired metaheuristic algorithm." Computers & Industrial Engineering 152 (2021): 107050.

[9]. Kalimuthan, C., and J. Arokia Renjit. "Review on intrusion detection using feature selection with machine learning techniques." Materials Today: Proceedings 33 (2020): 3794-3802.

[10]. Jyothsna, V. V. R. P. V., Rama Prasad, and K. Munivara Prasad. "A review of anomaly based intrusion detection systems." International Journal of Computer Applications 28, no. 7 (2011): 26-35.

[11]. Saxena, Aumreesh Ku, Sitesh Sinha, and Piyush Shukla. "General study of intrusion detection system and survey of agent based intrusion detection system." In 2017 International Conference on Computing, Communication and Automation (ICCCA), pp. 471-421. IEEE, 2017.

[12]. Jose, Shijoe, D. Malathi, Bharath Reddy, and Dorathi Jayaseeli. "A survey on anomaly based host intrusion detection system." In Journal of Physics: Conference Series, vol. 1000, no. 1, p. 012049. IOP Publishing, 2018.

[13]. Liao, Hung-Jen, Chun-Hung Richard Lin, Ying-Chih Lin, and Kuang-Yuan Tung. "Intrusion detection system: A comprehensive review." Journal of Network and Computer Applications 36, no. 1 (2013): 16-24.

[14]. Hajj, Suzan, Rayane El Sibai, Jacques Bou Abdo, Jacques Demerjian, Abdallah Makhoul, and Christophe Guyeux. "Anomaly-based intrusion detection systems: The requirements, methods, measurements, and datasets." Transactions on Emerging Telecommunications Technologies 32, no. 4 (2021): e4240.

[15]. Alamiedy, Taief Alaa, Mohammed Anbar, Zakaria NM Alqattan, and Qusay M. Alzubi. "Anomaly-based intrusion detection system using multi-objective grey wolf optimisation algorithm." Journal of Ambient Intelligence and Humanized Computing 11, no. 9 (2020): 3735-3756.

[16]. Keserwani, Pankaj Kumar, Mahesh Chandra Govil, Emmanuel S. Pilli, and Prajjval Govil. "A smart anomaly-based intrusion detection system for the Internet of Things (IoT) network using GWO–PSO–RF model." Journal of Reliable Intelligent Environments 7, no. 1 (2021): 3-21.

[17]. Dwivedi, Shubhra, Manu Vardhan, Sarsij Tripathi, and Alok Kumar Shukla. "Implementation of adaptive scheme in evolutionary technique for anomaly-based intrusion detection." Evolutionary Intelligence 13, no. 1 (2020): 103-117.

[18]. Dwivedi, Shubhra, Manu Vardhan, and Sarsij Tripathi. "Building an efficient intrusion detection system using grasshopper optimization algorithm for anomaly detection." Cluster Computing (2021): 1-20.

[19]. Shah, Ajay, Sophine Clachar, Manfred Minimair, and Davis Cook. "Building Multiclass Classification Baselines for Anomaly-based Network Intrusion Detection Systems." In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pp. 759-760. IEEE, 2020.

[20]. Ugtakhbayar, N., B. Usukhbayar, and S. Baigaltugs. "A Hybrid Model for Anomaly-Based Intrusion Detection System." In Advances in Intelligent Information Hiding and Multimedia Signal Processing, pp. 419-431. Springer, Singapore, 2020.

[21]. Tufan, Emrah, Cihangir Tezcan, and Cengiz Acartürk. "Anomaly-Based Intrusion Detection by Machine Learning: A Case Study on Probing Attacks to an Institutional Network." IEEE Access 9 (2021): 50078-50092.

[22]. Li, Kexin, Yong Zhang, and Shuai Wang. "An Intrusion Detection System based on PSO-GWO Hybrid Optimized Support Vector Machine." In 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1-7. IEEE, 2021.

[23]. Seraphim, B. Ida, E. Poovammal, Kadiyala Ramana, Natalia Kryvinska, and N. Penchalaiah. "A hybrid network intrusion detection using darwinian particle swarm optimization and stacked autoencoder hoeffding tree." Mathematical Biosciences and Engineering 18, no. 6 (2021): 8024-8044.

[24]. Mazini, Mehrnaz, Babak Shirazi, and Iraj Mahdavi. "Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms." Journal of King Saud University-Computer and Information Sciences 31, no. 4 (2019): 541-553.

7

Appendix

Appendix

Appendix (1) Source Code Proposed Method

```
//// main ////
clc
clear
close all
warning off
rng(0);
%% Loading Datset and preprocessing
global Xtr  Ytr Xts Yts
load("KDD_TRAIN.mat");
Xtr=y1(:,1:end-1);
Ytr=y1(:,end);
clear y1
load("KDD_TEST.mat");
Xts=y1(:,1:end-1);
Yts=y1(:,end);
clear y1
for i=1:size(Xtr,2)
Xtr(isnan(Xtr(:,i)),:)=mean(Xtr(:,i), 'omitnan');
end
for i=1:size(Xts,2)
Xts(isnan(Xts(:,i)),:)=mean(Xts(:,i), 'omitnan');
end
Xtr=double(Xtr);
Xts=double(Xts);
Ytr=double(Ytr);
Yts=double(Yts);
Xdata=normalize([Xtr;Xts],'range'); %min-max normalization
Xtr=Xdata(1:size(Xtr,1),:);
Xts=Xdata(size(Xtr,1)+1:end,:);
%% GWO-GEO for feature selection
Max_iteration=10; % Maximum number of iterations
SearchAgents_no=5;
dim=size(Xtr,2);
lb=0;
ub=1;
fobj=@FSCostFunction;
```

```
[Best_score,Best_pos,WOA_cg_curve]=GWO(dim,SearchAgents_no,
Max_iteration,lb,ub,fobj);
fs=find(round(Best_pos));
%% random forest for classification
Mdl = fitcensemble(Xtr(:,fs),Ytr);
[Trp scoresTrain] = predict(Mdl,Xtr(:,fs));
[Tsp scoresTrain] = predict(Mdl,Xts(:,fs));
cptr = classperf(Ytr,Trp);
cpts = classperf(Yts,Tsp);
disp(['Train accuracy=' num2str(cptr.CorrectRate)])
disp(['Train Precision='
num2str(cptr.PositivePredictiveValue)])
disp(['Train recall=' num2str(cptr.Sensitivity)])
fscore=2*((cptr.PositivePredictiveValue*cptr.Sensitivity)/(
cptr.PositivePredictiveValue+cptr.Sensitivity));
disp(['Train fscore=' num2str(fscore)])
disp(['Test accuracy=' num2str(cpts.CorrectRate)])
disp(['Test Precision='
num2str(cpts.PositivePredictiveValue)])
disp(['Test recall=' num2str(cpts.Sensitivity)])
fscore=2*((cpts.PositivePredictiveValue*cpts.Sensitivity)/(
cpts.PositivePredictiveValue+cpts.Sensitivity));
disp(['Test fscore=' num2str(fscore)])
```

//// **GEO** ////

```
function [x,xf] = GEO (fun,nvars,lb,ub,options)

%% initialization

PopulationSize = options.PopulationSize;
MaxIterations = options.MaxIterations;

ConvergenceCurve  = zeros (1, MaxIterations);

x = options.init; %lb + rand (PopulationSize,nvars) .* (ub-
lb);

FitnessScores = options.initF;

% solver-specific initialization
FlockMemoryF = FitnessScores;
FlockMemoryX = x;
```

```
AttackPropensity = linspace (options.AttackPropensity(1),
options.AttackPropensity(2), MaxIterations);
CruisePropensity = linspace (options.CruisePropensity(1),
options.CruisePropensity(2), MaxIterations);

%% main loop

for CurrentIteration = 1 : MaxIterations

    % prey selection (one-to-one mapping)
    DestinationEagle = randperm (PopulationSize)';

    % calculate AttackVectorInitial (Eq. 1 in paper)
    AttackVectorInitial = FlockMemoryX (DestinationEagle,:)
- x;

    % calculate Radius
    Radius = VecNorm (AttackVectorInitial, 2, 2);

    % determine converged and unconverged eagles
    ConvergedEagles = sum (Radius,2) == 0;
    UnconvergedEagles = ~ ConvergedEagles;

    % initialize CruiseVectorInitial
    CruiseVectorInitial = 2 .* rand (PopulationSize, nvars)
- 1; % [-1,1]

    % correct vectors for converged eagles
    AttackVectorInitial (ConvergedEagles, :) = 0;
    CruiseVectorInitial (ConvergedEagles, :) = 0;

    % determine constrained and free variables
    for i1 = 1 : PopulationSize
        if UnconvergedEagles (i1)
            vConstrained = false ([1, nvars]); % mask
            idx = datasample
(find(AttackVectorInitial(i1,:)), 1, 2);
            vConstrained (idx) = 1;
            vFree = ~vConstrained;
            CruiseVectorInitial (i1,idx) = -
sum(AttackVectorInitial(i1,vFree).*CruiseVectorInitial(i1,v
Free),2) ./ (AttackVectorInitial(i1,vConstrained)); % (Eq.
4 in paper)
        end
```

```
    end

    % calculate unit vectors
    AttackVectorUnit = AttackVectorInitial ./ VecNorm
(AttackVectorInitial, 2, 2);
    CruiseVectorUnit = CruiseVectorInitial ./ VecNorm
(CruiseVectorInitial, 2, 2);

    % correct vectors for converged eagles
    AttackVectorUnit(ConvergedEagles,:) = 0;
    CruiseVectorUnit(ConvergedEagles,:) = 0;

    % calculate movement vectors
    AttackVector = rand (PopulationSize, 1) .*
AttackPropensity(CurrentIteration) .* Radius .*
AttackVectorUnit; % (first term of Eq. 6 in paper)
    CruiseVector = rand (PopulationSize, 1) .*
CruisePropensity(CurrentIteration) .* Radius .*
CruiseVectorUnit; % (second term of Eq. 6 in paper)
    StepVector = AttackVector + CruiseVector;

    % calculate new x
    x = x + StepVector;

    % enforce bounds
    lbExtended = repmat (lb,[PopulationSize,1]);
    ubExtended = repmat (ub,[PopulationSize,1]);

    lbViolated = x < lbExtended;
    ubViolated = x > ubExtended;

    x (lbViolated) = lbExtended (lbViolated);
    x (ubViolated) = ubExtended (ubViolated);
    % calculate fitness
    for ii=1:PopulationSize
      FitnessScores(ii) = fun (x(ii,:));
    end

    % update memory
    UpdateMask = FitnessScores < FlockMemoryF;
    FlockMemoryF (UpdateMask) = FitnessScores (UpdateMask);
    FlockMemoryX (UpdateMask,:) = x (UpdateMask,:);

    % update convergence curve
```

```
ConvergenceCurve (CurrentIteration) = min (FlockMemoryF);
    fprintf ('iteration %d of GEO  \n', CurrentIteration);
end

%% return values
x = FlockMemoryX;
xf=FlockMemoryF;

///// GWO /////

% Grey Wolf Optimizer (GWO)
function
[Alpha_score,Alpha_pos,Convergence_curve]=GWO(dim,N,Max_ite
r,lb,ub,fobj)


lu = [lb .* ones(1, dim); ub .* ones(1, dim)];


% Initialize alpha, beta, and delta positions
Alpha_pos=zeros(1,dim);
Alpha_score=inf; %change this to -inf for maximization
problems

Beta_pos=zeros(1,dim);
Beta_score=inf; %change this to -inf for maximization
problems

Delta_pos=zeros(1,dim);
Delta_score=inf; %change this to -inf for maximization
problems

% Initialize the positions of wolves
Positions=initialization(N,dim,ub,lb);
Positions = boundConstraint (Positions, Positions, lu);

% Calculate objective function for each wolf
for i=1:size(Positions,1)
    Fit(i) = fobj(Positions(i,:));
end

% Personal best fitness and position obtained by each wolf
pBestScore = Fit;
```

```
pBest = Positions;

neighbor = zeros(N,N);
Convergence_curve=zeros(1,Max_iter);
iter = 0;% Loop counter

%% Main loop
while iter < Max_iter
    for i=1:size(Positions,1)
        fitness = Fit(i);

        % Update Alpha, Beta, and Delta
        if fitness<Alpha_score
            Alpha_score=fitness; % Update alpha
            Alpha_pos=Positions(i,:);
        end

        if fitness>Alpha_score && fitness<Beta_score
            Beta_score=fitness; % Update beta
            Beta_pos=Positions(i,:);
        end

        if fitness>Alpha_score && fitness>Beta_score &&
fitness<Delta_score
            Delta_score=fitness; % Update delta
            Delta_pos=Positions(i,:);
        end
    end

    %% Calculate the candiadate position Xi-GWO
    a=2-iter*((2)/Max_iter); % a decreases linearly from 2
to 0

    % Update the Position of search agents including omegas
    for i=1:size(Positions,1)
        for j=1:size(Positions,2)

            r1=rand(); % r1 is a random number in [0,1]
            r2=rand(); % r2 is a random number in [0,1]

            A1=2*a*r1-a;
% Equation (3.3)
            C1=2*r2;
% Equation (3.4)
```

```
            D_alpha=abs(C1*Alpha_pos(j)-Positions(i,j));
% Equation (3.5)-part 1
            X1=Alpha_pos(j)-A1*D_alpha;
% Equation (3.6)-part 1

            r1=rand();
            r2=rand();

            A2=2*a*r1-a;
% Equation (3.3)
            C2=2*r2;
% Equation (3.4)

            D_beta=abs(C2*Beta_pos(j)-Positions(i,j));
% Equation (3.5)-part 2
            X2=Beta_pos(j)-A2*D_beta;
% Equation (3.6)-part 2

            r1=rand();
            r2=rand();

            A3=2*a*r1-a;
% Equation (3.3)
            C3=2*r2;
% Equation (3.4)

            D_delta=abs(C3*Delta_pos(j)-Positions(i,j));
% Equation (3.5)-part 3
            X3=Delta_pos(j)-A3*D_delta;
% Equation (3.5)-part 3

            X_GWO(i,j)=(X1+X2+X3)/3;
% Equation (3.7)

        end
        X_GWO(i,:) = boundConstraint(X_GWO(i,:),
Positions(i,:), lu);
        Fit_GWO(i) = fobj(X_GWO(i,:));
    end
    options.PopulationSize = N;
    options.MaxIterations  = 10;
    options.AttackPropensity = [0.5 ,   2];
    options.CruisePropensity = [1   , 0.5];
```

```
    options.init=X_GWO;
    options.initF=Fit_GWO;
    [xbestGEO,xFGEO] = GEO
(fobj,dim,zeros(1,dim),ones(1,dim),options);
    X_GWO = xbestGEO;
    Fit_GWO=xFGEO;
    %% Calculate the candiadate position Xi-DLH
    radius = pdist2(Positions, X_GWO, 'euclidean');
% Equation (10)
    dist_Position = squareform(pdist(Positions));
    r1 = randperm(N,N);

    for t=1:N
        neighbor(t,:) = (dist_Position(t,:)<=radius(t,t));
        [~,Idx] = find(neighbor(t,:)==1);
% Equation (11)
        random_Idx_neighbor = randi(size(Idx,2),1,dim);

        for d=1:dim
            X_DLH(t,d) = Positions(t,d) + rand
.*(Positions(Idx(random_Idx_neighbor(d)),d)...
                - Positions(r1(t),d));
% Equation (12)
        end
        X_DLH(t,:) = boundConstraint(X_DLH(t,:),
Positions(t,:), lu);
        Fit_DLH(t) = fobj(X_DLH(t,:));
    end


    %% Selection
    tmp = Fit_GWO < Fit_DLH;
% Equation (13)
    tmp_rep = repmat(tmp',1,dim);

    tmpFit = tmp .* Fit_GWO + (1-tmp) .* Fit_DLH;
    tmpPositions = tmp_rep .* X_GWO + (1-tmp_rep) .* X_DLH;

    %% Updating
    tmp = pBestScore <= tmpFit;
% Equation (13)
    tmp_rep = repmat(tmp',1,dim);

    pBestScore = tmp .* pBestScore + (1-tmp) .* tmpFit;
```

```
    pBest = tmp_rep .* pBest + (1-tmp_rep) .* tmpPositions;

    Fit = pBestScore;
    Positions = pBest;

    %%
    iter = iter+1;
    neighbor = zeros(N,N);
    Convergence_curve(iter) = Alpha_score;
    disp(['iter: ' num2str(iter) 'Best Fitness:'
num2str(Alpha_score)]);
end
end
```