Turkish Journal of Computer and Mathematics Education Vol.12 No. 5 (2021), 1226-1236

Research Article

A Hybrid Adaptive Development Algorithm and Machine Learning Based Method for Intrusion Detection and Prevention System

K. NandhaKumar[a], Dr.S. Sukumaran[b]

[a]Ph.D Scholar, Department of Computer Science Erode Arts and Science College, Erode, Tamilnadu, India. E-mail: nandha.k07@gmail.com
[b]Associate Professor, Department of Computer Science, Erode Arts and Science College, Erode, Tamilnadu, India. E-mail: prof_sukumar@yahoo.co.in

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: Network Intrusion detection and prevention Systems (NIDPS) are employed in monitoring a network which safeguards user integrity, privacy thereby ensuring the data security and availability in a network. Such systems not only monitor the suspicious activities in a network but also used as control systems to eliminate the malicious users from the network. In this paper, a Hybrid Adaptive Development Algorithm and Machine Learning Algorithm (ADA-MLA) method is proposed to identify the malicious activities and eliminating them from the network. The deployment of honeypot-based intrusion is improved adaptive development algorithm. Machine learning algorithm has been employed in the Hybrid IDPS for learning the network data patterns which also identifies the maximum probable attacks in the network. The signatures for the DARPA 99 data set have been updated during the implementation of intrusion prevention system on a real-time basis. The hybrid method works on (i) classifying the attacks based on protocols and (ii) classifying the attacks on pre-determined threshold values. Hence, both known and unknown attacks can be easily captured in the proposed hybrid IDPS method which thereby achieves higher attack detection and prevention accuracy while compared to the conventional attack detection and prevention methodologies.

Keywords: Machine Learning, Honeypot, Improved Adaptive Deployment, Attack, Detection Rate, Learning and Decision.

1. Introduction

Though the communication and information technologies are rapidly developing, the intrusions over the systems are simultaneously increasing. Hence, prevention of systems from various attacks is significant to ensure information security in order to satisfy the needs of people. Several researches and studies were carried out by the research community to overcome the anomalous intrusions through various prevention and detection method. An intrusion prevention system (IPS) regularly monitors a network and traces malicious intrusions and gathers information about those intrusions. IPS secures the network by identifying and by preventing attacks. Any incident which attempts to violate the pre-defined security practices will be identified by the IPS. Defining a honeypot is complex because of its constantly varying nature but can be utilized in several security aspects including attack detection, prevention of the identified attack and gathering information from the identified attack. Honeypot is a growing technology, distinct and more generalized and therefore utilizing this technology to address the security issues is difficult. But, it can be more helpful in intrusion prevention and detection and is widely used to detect network vulnerabilities. Honeypot can be considered as a network security tool where the merit of the tool depends on whether it is capable of probe, attack or compromise. The construction of the signatures for network intrusion detection systems is quite complex since a complete understanding of the traffic features of the characteristics are required to be identified by the novel signature presented. If the signatures are declared simpler, it could produce higher amounts of false positives and majority of the specific items could result in false negatives. These issues in honeypots are of major concern which needs to be addressed while utilizing them for the intrusion prevention and detection systems.

The research community adds several functions to the IDS but malicious users find ways to detect, bypass and disable the system before they intrude to the framework which results in the denial of service (DoS) attack. Hence, security engineers employ the architecture of intrusion detection and prevention systems (IDPS) to overcome these attacks. IDPS is invisible to the malicious intrusions that restrict the communication allowed between several security elements present in a network. The increasing number of threats requires swift detection of attacks for which several studies are made on the IDPS in the literature. Generally, the design of the IDPS are built to identify the malicious attacks that are categorized into three types namely machine learning, knowledge-based and statistical approaches. The machine learning approaches are extensively used for constructing the IDPS because of their proven results in the detection and prevention of attacks. This method works well in avoiding the system damage from both known and unknown attacks.

2. Literature Survey

The related work of this paper can be divided into three classes such as (i) machine learning algorithms related to this study (ii) the DARPA 99 dataset used in the study and (iii) the honeypot techniques related to the IDPS.

Currently, the machine learning algorithms are extensively employed in various researches related to the network intrusion detection systems. Since various datasets are made available opensource, several methods utilized these datasets to overcome the issues concerned in the detection and prevention of threats. In most cases, the machine learning approaches are combined with the stand-alone data analytics methods as hybrid methods which make use of layered as well as hierarchical models [1,2], detection of anomalies [3] and enhancing the machine learning approaches by including the knowledge-based methods [4]. Majority of these approaches primarily aim to detect rare attacks without compromising the accuracy of detection of the recurrent attacks and reducing the false alarm rate in those detected attacks. Various machine learning approaches such as SVM, fuzzy logic approach and neural network-based approaches are presented in intrusion detection. Ahmed et.al [5] introduced an IDS by combining various classifying methods based on rule based and decision tree approaches such as J-Rip, REP tree and Forest PA. Among these approaches, the first two considers the input characteristics of the dataset and classifies the traffic of the networks as benign or attack. The Forest PA approach employs the initial dataset features along with the first classifier's output and the inputs of the second one. Authors made use of CICIDS-2017 dataset and achieved 97% accuracy where the rate of detection being 95%. This hybrid approach gained better results with this dataset while compared to other methods. Various studies utilized hybrid approaches rather than the single approaches [6]. A hybrid layered approach which combines multiple classifiers including naive bayes tree (NB-Tree) and naïve bayes classifier (NBC) is presented by Sharma et.al [7] for enhancing the recall (attack detection ratio) and accuracy of minority class including U2R and R2L attacks without affecting the majority class performance. This approach combines the precision and recall for both vital and minor threats and retains the false positives at bearable levels in the IDS. This system employs contrasting reduced feature sets for predicting every class of attack through periodically conducted experiments and domain knowledge. This approach managed to achieve 99.05% total accuracy and showed better results in classifying minor classes while comparing to the standard methods. Multi-layer perceptron (MLPs) is a better approach in identifying normal attacks and link attacks [8]. Moradi and Zulkernine [9] developed a neural network-based method to detect intrusions. Authors employed MLP to detect intrusions on the basis of off-line evaluation method. This approach is aimed to address a multi-class issue where it detects the attack type as well.Various structures of neural networks are evaluated to identify the best neural network considering the number of hidden layers. They also utilized a preliminary prevention validation approach in the training for improving the neural network's rationalization ability. The experimental results classified the logs with around 91% of accuracy considering the neural network with two number of hidden neuron layers and it achieved to about 87% of accuracy considering only one hidden layer. Though this approach solved a three-class issue, it fails to cover various classes. Zhang et.al [10] analysed the need for the artificial neural network in detecting intrusions. The neural networks manage noisy data quite effectively but they need higher amount of data to train them. Further, it is very difficult to choose the optimal and practicable framework for a neural network. Gupta et.al [11] developed a standard and easier layered approach to detect intrusions. They aimed to present an approach with higher amount of accuracy and computationally intensive. Their approach is of three layers including data integrity, confidentiality and integrity where every layer in the system refers to a security aspect. The initial layer is responsible for establishing connectivity and manages the features at the packet level including the number of host connections, addresses of source and destination, ID of the user, port numbers of source and destination layers respectively. This layer is further optimized to identify the attacks concerned to the aspects of availability including probes, U2R, R2L and DoS attack. The data privacy layer is responsible for data confidentiality and contains some of the features including the amount of data retrieved, number of files accessed etc., The access control layer is responsible for data integrity and it manages privileges to the user, modifications in files etc., Ibrahim et.al [12] considered the issue of performance optimization in the IDS and introduced a multi-layer model to detect intrusions. Authors employed various machine learning approaches such as Naïve Bayes, C5 decision tree approach and MLP neural networks by utilizing the gain ratio to choose the optimal features for every layer in order to utilize the limited space for storage and to achieve better performance in detecting intrusions. The implementation results proved C5 detection tree approach along with the presented multiple layer model achieved improved accuracy ratio in classification with the selected features using gain ratio and lesser false alarm rate while compared to Naïve Bayes and the MLP neural networks approach. But this system is capable of detecting limited number of intrusions.

Many of the intrusion detection systems utilized NSL-KDD, DARPA99 and KDD-CUP99 datasets. Such datasets make use of various feature selection methods to function. We shall discuss some of those related studies in this section. Since many of the feature selection algorithms fail to achieve the optimal features more efficiently as well as effectively, Wang and Feng [13] introduced a novel hybrid algorithm for feature selection by assessing the pitfalls present in several existing relevant measurements of feature selection. They introduced the feature selection approach by pre-processing the samples and two feature subsets which are fetched through two contrasting filters. Then, they present a union function based on feature weights for merging the fetched feature sub-sets. Since the clustering protocol is capable of achieving best quality clusters even without the need of a cluster member, they introduced it to get the conclusive feature sub-set by employing a fixed threshold value. Authors used eight data sets including KDD CUP-99 in two important classifiers such as KNN and SVM. This method achieved higher accuracy in classification with enhanced speed of execution but it is not much efficient in searching the thparameter. Saurabh and Sharma presented a reduction approach based on feature vitality to find the significant diminished input characteristics. Authors introduced Naïve Bayes approach on the diminished datasets to detect intrusions. Though this approach had achieved better accuracy in classification but it is not so effective in detecting the U2R attacks with limited overheads and decreased complexity. Yousefi et.al considered the issue of higher dimensionality and presented a feature selection approach with two algorithms for feature selection and evaluated the efficacy of the proposed protocol while compared to that of feature selection approach based on mutual information. The selected approach employs a feature selection technique on the basis of the measure of feature goodness where linear as well as non-linear measures such as mutual information and linear co-efficient of correlation are considered. Least squares based SVM technique along with machine learning approach is used in the IDS and the results were assessed using KDD CUP-99 dataset. The results proved that the proposed method achieved improved accuracy in intrusion detection specifically in the U2R and R2L attacks. Akashdeep et.al present a smart intrusion detection system that initially operated the feature ranking depending on the correlation and gain. Further, feature reduction is accomplished by merging the ranks which are fetched from the correlation and information gain through the proposed method to find the relevant and irrelevant features. Such diminished features are further forwarded to a feed forward NN to train and test them in the KDD CUP-99 dataset. The proposed approach smartly classifies the test data into the class of attack and non-attack. Authors also utilized five more datasets and compared the proposed approach with several performance metrics. Though the system achieved appreciable results, increasing the data quantity needs high power networks. Ayman et.al developed a novel feature selection method to efficiently choose the significant required features to detect intrusions. Since discarding the unsignificant and unnecessary features aids to construct a swift process to train and test which will also decrease the utilization of the resources by managing the larger rate of detection. The performance of the system is evaluated using the KDD dataset which achieved the higher rate of detection with faster detection ratio. Fengli et.al introduced a feature selection method on the basis of Bayesian network classification using NSL-KDD dataset. The results depicted that the features chosen by the proposed method had reduced the attack detection time and improved the accuracy of classification along with the ratio of true positives. This approach is not so effective in detecting the U2R attacks. Muhammad and Dewanintroduced a novel method that integrates the selection of features and classification for several class DARPA 99 data set to detect intrusions by utilizing the SVM approach. This approach also utilized limited amount of training features and avoids unnecessary feature sub-sets. The proposed approach was evaluated and found classification accuracy of about 91% utilizing three number of features and achieved 99% of accuracy in classification by employing 36 number of features and the total 41 features for training had achieved 99% of the accuracy in classification.

In the recent days, the detection of malicious intrusions grabbed extensive attention because of which safeguarding the privacy of sensitive data in networked systems is challenging. Though a variety of security approaches are proposed, they still carry few restrictions. Majority of the current research studies focus in machine learning approaches to detect intrusions through the collection of data using several information technologies namely honeypots. Owezarski et.al [14] presented a self-governing approach to detect attacks through unsupervised anomaly learning approach by utilizing the information gathered by the designed honeypots. Authors employed cluster-based approaches including sub-space clustering approach, accumulation of evidences and density conscious clustering to classify the flow entities during variety of traffic. The primary benefit of this approach is that it avoids the need of training phase. Luo et.al developed a smart honeypot-based model to enhance the security of IoT based devices using machine learning approaches. In this approach, the designed honeypot is used to manage the intrusions through model optimization. Lee et.al [15] developed a model that automatically categorizes the social spams using machine learning approaches which can be utilized in a variety of social media platforms including MySpace, Facebook etc., through social honeypots to collect data about the anomalous intrusions. Utilizing the link defense system proposed by Feng et.al [16], Li.et.al introduced a honeypot model to address the issues present in the conventional tools. The link approach proposed by the authors improves the communication and management among the defense model components and the designed honeypot using SNMP protocol. To avoid new attacks, a central honeypot guards the defense system

and decides whether to allow or block depending on the honeypot state. Kamel et.al [17] developed a honeypot model by utilizing machine learning algorithms to construct an intelligent agent which prevents cyber-attacks by predicting them. Inspired by this approach a hybrid approach is introduced in this study that makes use of machine learning algorithms and honeypot model to prevent intrusions. The objective of this research study is to develop an intrusion prevention and detection system through honeypot model.

Machine learning approach belongs to the area of artificial intelligence which aimed to avoid the conventional programming methodologies by providing an opportunity to the computer-based systems to learn by understanding the methods, data structure and construct them into models which are then utilized to solve various complex problems [18]. Machine learning models operates on the basis of mathematical hypothesis functions where the hypothesis function clearly plots from the variables of the inputs to the outputs. They are also used to identify the clusters or structures present in them. Machine learning approaches can be classified into supervised learning, un-supervised learning and re-inforcement learning methods. The supervised learning approach contains prediction function that helps to learn from examples that are labelled or from the input/output database pairs. They can be categorized into classification and regression problems [19]. The classification problems let the model to output the results within the discrete set of values whereas the regression problems allow the model to output the results within the continuous set of values. Support vector machines (SVMs), decision tree models (DTMs) linear regression models (LRMs) are the widely used supervised algorithms. In contrast, the un-supervised learning approach contain data with zero labels which produces coarse observations of arbitrary variables. Semi-supervised learning is another approach which contains both data: labelled and unlabelled. This approach enables to assess huge amount of data even without any labels. Re-inforcement learning differs from the supervised learning that contains no labels at all but the agent is capable of learning to accomplish the allotted task [20]. As a part of this study, we employ decision trees, k-means clustering algorithm and linear regression as machine learning approaches. We shall discuss these topics hereunder.

KM is a widely applied cluster-based algorithm and is an algorithm that operates iteratively. KM algorithm initialize the centroids in the cluster arbitrarily and move towards the centroids till each point is allotted to the adjacent cluster which is nearer by making the centroids smaller. The sum of squared Euclidean distances among the centroid C_k and every data point ($DP_1, DP_2 \ldots\ldots DP_N$) which can be termed as the cluster error CE.

$$CE(C_1, C_2 \ldots\ldots C_P) = \sum_1^N \sum_1^P ||DP_{i-} C_P||^2 \quad (1)$$

The KM clustering algorithm provides localized best solutions concerned to the cluster error. Though the algorithm is extensively used in a variety of applications, it lacks in the sensitivity to the number of clusters and the primary positions of the cluster centres [21]. Hence, to get the nearest best solutions by deploying the KM algorithm, various trials has to be scheduled by varying the cluster centres initial positions. Global KM algorithm is proposed to resolve this issue.

LRP is generally utilized to compute a function of linear hypothesis among the input and the output variables as a tool for classification and regression and can be stated as:

$$hf_{\phi}(S) = \phi_0 + \phi_1 . S_1 + \ldots\ldots\ldots + \phi_n . S_n \quad (2)$$

where hf_{ϕ} is referred to as the function of hypothesis,ϕ_n denotes the hypothesis function weights andS_i denotes the input variables. To compute the values of weights, initially the error among the anticipated result (X) and the approximated result (X^2), cost function called mean square error is used which can be written as:

$$R = \frac{1}{K} \sum_{n=1}^{K} \left(X^{a^{(n)}} - b^{(n)} \right)^2 \quad (3)$$

Further, gradient descent algorithm is implemented [22] which is regarded as the significant linear regression step. It enables to compute the best weight values by reducing the cost function. The algorithm is iterative that updates the weights at every iteration to reduce the cost function by determining a value of threshold.

Decision tree is a method of representing the data structures hierarchically through tests or decision sequences to predict the results. Considering various characteristics, the decision will be initiated by considering any of the characteristic and the process is repeated till an optimal characteristic is matched. Every observation will be allotted to a class that contain a set of variables which is tested in the nodes of the tree. The internal nodes are used to evaluate the experiments and the leaf nodes are used to perform decisions [23]. This approach is suited to solve regression as well as classification problems.

Generally, the attacks will depend on several tools which aims to scan the overall network for intrusions. The main objective of honeypots lies in letting the intruder to trust that he is capable of taking over the entire system that enables the administrator to track the reason in compromising the intruders to protect them against newly discovered attacks and let the attacker to take ample amount of time to react [24].The implementation of honeypots is highly flexible and manages several forms. Majority of the research studies classifies honeypots depending on their usefulness or the interactions they can support [25]. In that basis, honeypots are classified as low-interaction, medium interaction and high interaction honeypots. Some studies classify honeypots on the basis of how they are utilized. In this category, they are classified as research-oriented and production-oriented honeypots. In case of low interaction honeypots [26], they are constrained to the emulation degree supported by the honeypot and therefore, the communication among the honeypot model and the intruder will be very lesser. This type of honeypot provides some benefits to the attacker but he will have only lesser amount of scope. For an instance, if the honeypot emulates the service of file transfer protocol on the port number 56, it is capable of emulating either the log-in command or some other command. Honeypots offer several benefits including their easy implementation, simplicity and they are not too risky. Low interaction honeypots document less amount of information and tracks down only known attacks. Further, the emulated services from the honeypot is not capable of offering much. Hence, an intruder can easily identify this type of honeypot. Some examples of low interaction honeypots are LaBrea, GHH, Tiny honeypot, deception toolkit etc., Some studies concentrated on Medium interaction honeypots [27] that enables the attacker to grant further more access while compared to that of low interaction honeypots. Though it allows to log several advanced attacks and provides enhanced simulation services, they need large amount of time for implementation and also require specialization. High interaction honeypots are also a class of honeypots [28] that requires the deployment of real-world applications and real-world operating systems. But it is too risky because the overall network is compromised. Honeynet is an example of high interaction honeypot.

In this study, we utilize an improved adaptive honeypot deployment algorithm with machine learning to enhance the network security. Honeypot deployment will be based on the decoy network which monitors the internal or external threats in a network system. Therefore, it has to be installed after the firewall or in a de-militarized zone (DMZ) or before the firewall [54]. The primary benefit of selecting the initial position (as in Fig.1) is that variations are not made in the filtering rules of the firewall that safeguards the internal network. Further, there will not be any recent risks for the machines present in the internal network because of the incorporation of the location rather it will not identify any threats within the network. Usually, the firewall blocks all the outgoing flows in the network. Next,
the honeypot will be placed in a DMZ.

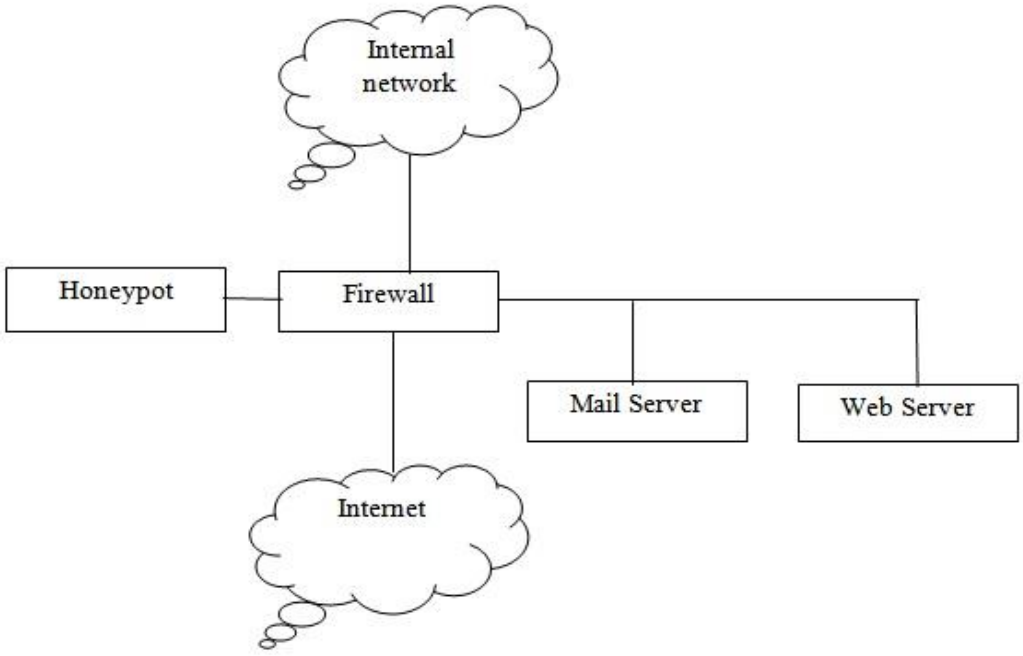

Figure 1. Deployment of Honeypot

The benefit of utilizing the honeypot in the DMZ is that it allows the public servers in the internet that isolates from the internally present network. The DMZ (Fig 2) can be utilized in the production servers or is

committed to a honeypot. Considering a firewall, it lets only the incoming traffic to allow them to the DMZ for the accessible services. This allows the DMZ to evaluate the aimed attacks alone for the questionable services.

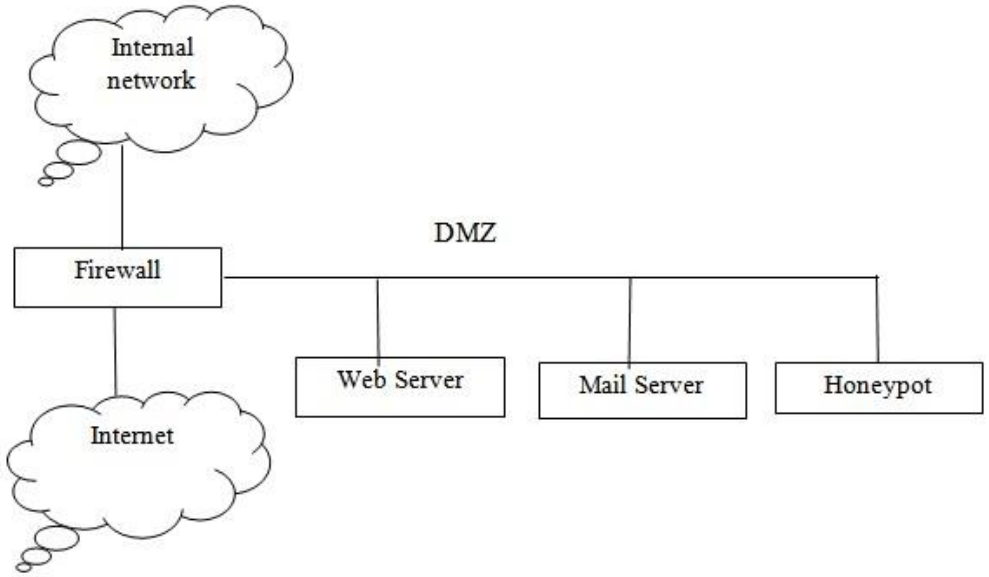

Figure 2. Deployment of Honeypot in DMZ Using Improved Adaptive Algorithm

Honeypot can also be deployed in LAN. Considering a decoy model, honeypots identify external threats and also persuade higher risks of susceptibilities. Once it gets compromised, the decoy model can be utilized by the intruder to introduce some other threats in the internally present network. This position also lets identifies threats from users within the organization to the inbound services or identifying a poor configuration of firewall [54]. Some of the tools that can be used to setup honeypots are Specter, Netfacade, CurrPorts and KF Sensor.

3. Proposed Work

We present a solution by combining the improved adaptive honeypot deployment algorithm and machine learning algorithm (ADA-MLA) method is capable of collecting data, analyse them and predict the intrusions to assure the security of a network. The overall implementation diagram of the proposed system is discussed in detail in the Figure 3.

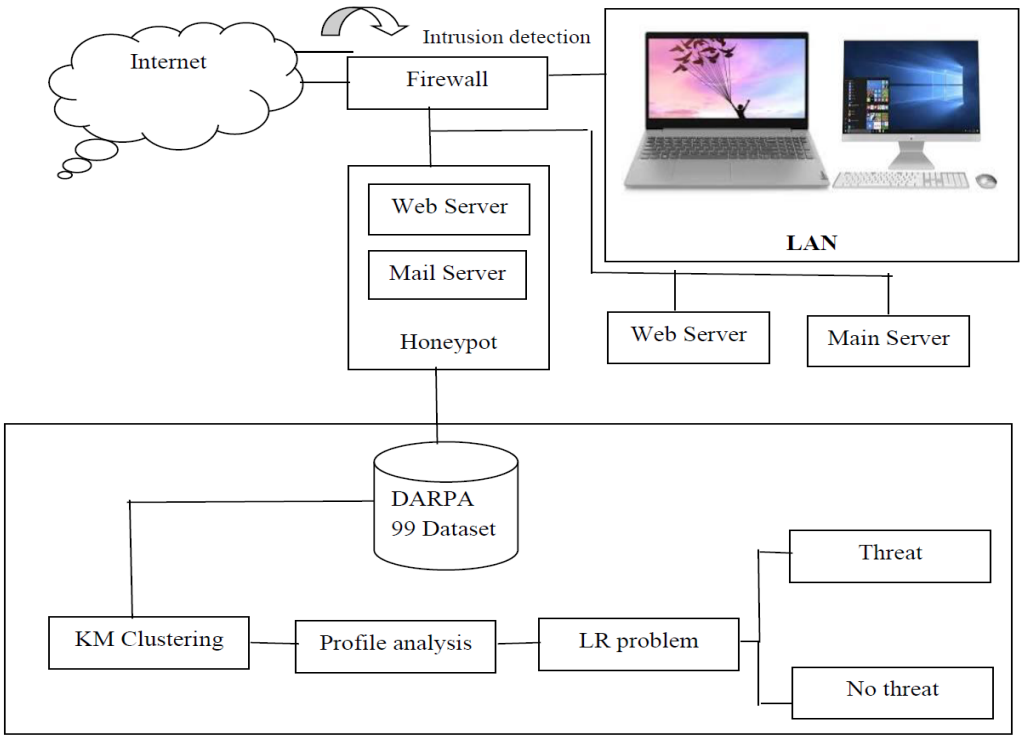

Figure 3. Overall Implementation of the Proposed Method

The proposed method enables to detect anomalous pattern of profiles over the offered services through analysis using machine learning approach. The machine learning predicts intruder profiles and consider the decisions by re-configuring the firewall to stop the threats from invasion. The configuration of the firewall is to be done in a way to re-orient the anomalous flow to the designed honeypot model to collect data regarding the layers including transport and application respectively. Then, the gathered information will be passed to the hybrid algorithm. KM clustering algorithm clusters the data into analogous categories and builds the profile to the user. The constructed profile is then categorized as a threat or no threat depending on the classifying algorithm.

During every iteration that takes place with anomalous traffic, the designed honeypot will save the gathered data vector including the log, internet protocol (IP) address and length of the packet $Vect_n^{user}$ in the DARPA 99 dataset to build the profile of the user. Further, KM clustering algorithm clusters the data into analogous categories to build the user profile. LR problem is employed to model every class to provide a highly important and analogous data presentation as discussed in the Figure 4 and 5 which details about two stages namely learning and decision process. Considering a vector $Vect_n^{user}$ with i number of elements of data gathered by the designed honeypot model where the $Vect_n^{user}$ information will convert them to subjective data I_{sub} and perceptible data I_{per}.

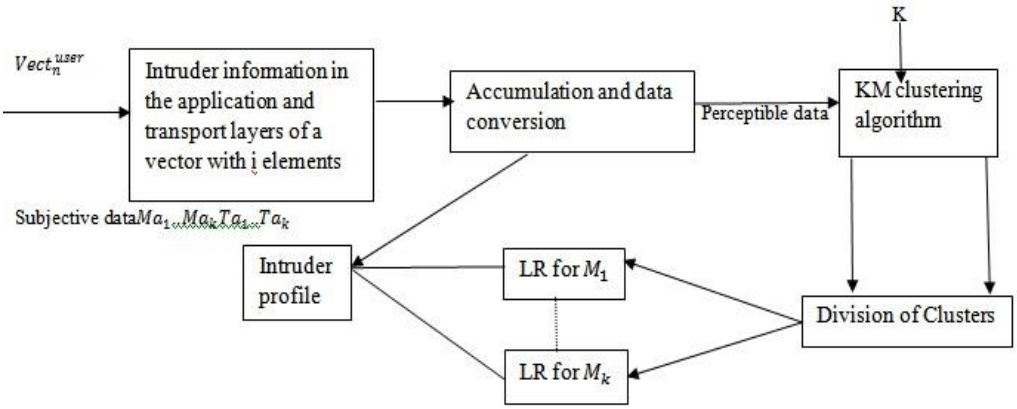

Figure 4. Learning Process

The subjective data includes the IP address which is stored in the profile of the user and the objective data such as packets inter-time will be clustered into analogous groups through KM clustering algorithm, M_j[K]. Every generated class will be denoted using the linear regression model, ML_j[r]. The subjective data along with the presented perceptive design generates the anomalous user profile. Distance metric is used to take the decision by considering the anomalous user profile. The distance metric will be computed among the anomalous user profile and the training models present in the learning process.

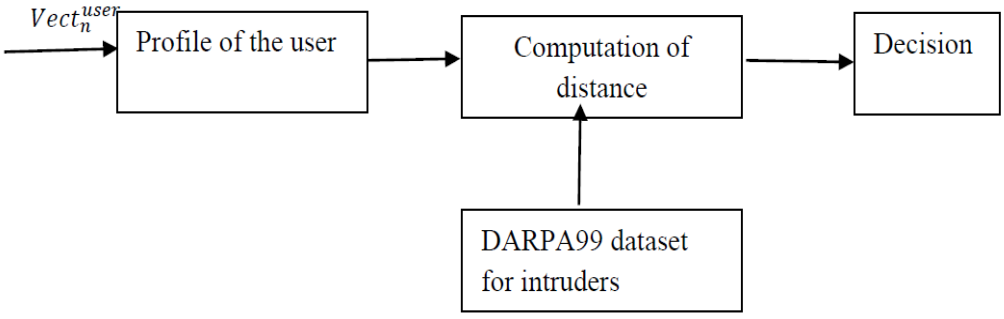

Figure 5. Decision Process

The learning process shown in the Fig 4 is represented in the algorithm1 that requires three inputs such as the required number of clusters, centroid initialization and the LR parameters. The accuracy of the initialization process will be higher and the estimation of the LR parameters, the fitting model accuracy will be higher.The algorithm2 returns the classification and the creation of user profiles on the basis of the model created for attackers.

Algorithm 1: Learning Process

Input: n clusters

$Vect_{pk} = (Vect_1^{user}, Vect_2^{user} \ldots\ldots Vect_k^{user})$ // vector array denotes the information of intruders

BEGIN

$D_{ij} = Qual[Vect_{pk}]$ // Qualitative transformation

$E_{xy} = Quan[Vect_{pk}]$ // Quantitative transformation

$h = m - 1$ //LR order = dimension of space-1

$for\ b = 1; b < m, b++$ //for every row of I_{sub}

M_j[K], S_j[K] = KM(E_{iy},K) // the radius of the constructed cluster center

$for\ b = 1; b < m, b++$

ML_j[K]= LR(M_j[K], S_j[K],I_{sub}) //Each cluster is applied with linear regression

OUTPUT // profile of the intruder

D_{ij} // Qualitative transformation

ML_m[K] //co-efficient of linear regression

Algorithm 2: Decision Process

Input: n clusters

$Vect_{pk} = (Vect_1^{user}, Vect_2^{user} \ldots\ldots Vect_k^{user})$ // vector array denotes the information of intruders

$(MP_{ij}, ML_m$[K]) //profiles of intruders

BEGIN

$UD_{ij} = Qual[Vect_{pk}]$ // Qualitative transformation

$E_{xy} = Quan[Vect_{pk}]$ // Quantitative transformation

$h = m - 1$ //LR order = dimension of space-1

$for\ b = 1; b < m, b++$ //for every row of I_{sub}

M_j[K], S_j[K] = KM(E_{iy},K) // the radius of the constructed cluster center

$for\ b = 1; b < m, b++$

UML_j[K]= LR(M_j[K], S_j[K],I_{sub}) //Each cluster is applied with linear regression

OUTPUT // profile of the intruder

$Dist(ML_m$[K], UML_j[K])

Is_equal(UD_{ij},MD_{ij})

This algorithm is for the process of decision. ML_m[K]in Euclidean is the nearest vector to ML_m[K]where the decision accomplished will depend on the latest profile projection on the profiles of intruders. If the minimum distance value fails to surpass the threshold value, the same will be moved to the class of minimal distance and construct a latest cluster centre. Else, yet another class will be created using ones and the procedure will be repeated.

The data will be classified using the process of protocol analysis through which the data will be downloaded for evaluation. Misuse detection approach is adopted to improve the accuracy of attack detection. The rules for intrusion will be dynamically included using honeypot to improve the pattern matching approach performance.

4. Simulation Results

In this approach, we utilized the DARPA 99 data set to evaluate the intrusion detection system using honeypot. In this experiment, we utilize 10000 records with normal behaviour and 1000 records with

abnormality. We use several number of thresholds and is the value of threshold is 40, the rate of detection of probe attacks and U2R attack will be larger and the rate of mission will be lesser. If the value of threshold is lesser, the regular logging will be remote from the regular cluster which gives rise to a larger rate of mission. If the value of threshold is 50, the generated results will be lesser. The simulation results are compared with the conventional K-means clustering with that of proposed hybrid approach. The performance analysis of both of these approaches are evaluated with respect to the performance metrics false positive rate, precision and recall. The following tables depict the comparison chart of the existing K-means clustering and proposed hybrid approach.

Table 1. Results of Simulation Using Various Threshold Values – K Means Clustering

Value of Threshold	Type of Intrusion			
	Probe attack	DoS attack	U2R attack	R2L attack
10	79.98	78.86	79.32	80.32
20	77.34	76.54	77.12	77.67
30	75.89	76.11	75.56	75.10
40	74.98	74.47	74.12	74.10
50	74.32	73.98	73.10	73.87

Table 2. Performance Analysis of Various Classes of Attacks Using K Means Clustering

Type of Attack	Precision	Recall	False Positive Rate
Probe	0.98	0.98	0.027
DoS	0.99	0.99	0.145
U2R	0.97	0.96	0.034
R2L	0.97	0.97	0.056
Normal	0.99	0.99	0.154

Table 3. Results of Simulation using Various Threshold Values – Proposed Method

Value of Threshold	Type of Intrusion			
	Probe attack	DoS attack	U2R attack	R2L attack
10	99.14	99.29	99.45	99.34
20	87.56	90.34	90.23	90.78
30	85.23	89.98	88.34	89.34
40	83.24	88.22	86.45	87.97
50	82.31	86.34	85.38	86.28

Table 4. Performance Analysis of Various Class of Attacks using Proposed Method

Type of Attack	Precision	Recall	False Positive Rate
Probe	0.45	0.41	0.012
DoS	0.34	0.33	0.011
U2R	0.31	0.30	0.009
R2L	0.29	0.28	0.007
Normal	0.11	0.10	0.011

5. Conclusion

In this study, we proposed a hybrid method that combines the machine learning techniques and the honeypot construction using improved adaptive deployment algorithm (ADA-MLA) to ensure the network security against both known and unknown attacks. The proposed method is better and efficient enough since it contains two significant information where one information is used to create profiles and the other is used to classify the created profiles. The proposed hybrid method creates a solid model and emerge as a predictive model that effectively recognizes anomalous intrusions and classifies the threats. Thus, the proposed system is modelled as intrusion detection and prevention system that effectively recognizes the class of attacks and stops those attacks from invading over the secure network. Honeypot is used to gather large number of threats and databases of the prevailing rule set are updated periodically. DARPA 99 dataset is used by the machine learning and honeypot deployment using improved adaptive algorithm improves the rate of detection. This study improves the performance and safety of the overall network and can be utilized in large scale networks as well.

References

Ahmim, A., & Zine, N.G. (2015). A new hierarchical intrusion detection system based on a binary tree of classifiers. *Information & Computer Security*, *23*, 31–57.

Kevric, J., Jukic, S., & Subasi, A. (2017). An effective combining classifier approach using tree algorithms for network intrusion detection. *Neural Computing and Applications*, *28*(1), 1051-1058.

Aljawarneh, S., Aldwairi, M., & Yassein, M.B. (2018). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. *Journal of Computational Science*, *25*, 152-160.

runadevi, M., & Perumal, S.K. (2016). Ontology based approach for network security. *In Proceedings of the International Conference on Advanced Communication Control and Computing Technologie*s (ICACCCT), Ramanathapuram, India, *25–27*, 573–578.

Ahmim, A., Maglaras, L., Ferrag, M.A., Derdour, M., & Janicke, H. (2019). A novel hierarchical intrusion detection system based on decision tree and rules-based models. *In 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 228-233.

Guo, C., Ping, Y., Liu, N., & Luo, S.S. (2016). A two-level hybrid approach for intrusion detection. *Neurocomputing*, *214*, 391-400.

Sharma, N., & Mukherjee, S. (2012). A novel multi-classifier layered approach to improve minority attack detection in IDS. *Procedia Technology*, *6*, 913-921.

Cannady, J. (1998). Artificial neural networks for misuse detection, *Proceedings of the National Information Systems Security Conference* (NISSC'98), Arlington, VA, 443-456.

Moradi, M., & Zulkernine, M. (2004). A neural network based system for intrusion detection and classification of attacks. *In Proceedings of the IEEE international conference on advances in intelligent systems-theory and applications* 15-18.

Debar, H., Becker, M., & Siboni, D. (1992). A neural network component for an intrusion detection system. *In Proceedings 1992 IEEE Computer Society Symposium on Research in Security and Privacy*, IEEE, 240-250.

Gupta, K.K., Nath, B., & Kotagiri, R. (2006). Network Security Framework, Int'l *International Journal of Network Security*, *6*(7), 151-157.

Ibrahim, H.E., Badr, S.M., & Shaheen, M.A. (2012). *Adaptive layered approach using machine learning techniques with gain ratio for intrusion detection systems. arXiv preprint arXiv:1210.7650*.

Wang, Y., & Feng, L. (2018). Hybrid feature selection using component co-occurrence based feature relevance measurement. *Expert Systems with Applications*, *102*, 83-99.

Owezarski, P. (2014). Unsupervised classification and characterization of honeypot attacks. *In 10th International Conference on Network and Service Management (CNSM) and Workshop*, IEEE, 10-18.

Lee, K., Caverlee, J., & Webb, S. (2010). Uncovering social spammers: social honeypots+ machine learning. *In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 435-442.

Feng, G., Zhang, C., & Zhang, Q. (2013). A design of linkage security defense system based on honeypot. *In International Conference on Trustworthy Computing and Services*, 70-77.

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing," *EURASIP Journalon Advances in Signal Processing*, *2016*(1), 67.

Ullah, Z., Al-Turjman, F., Mostarda, L., Gagliardi, R. (2020). Applications of artificial intelligence and machine learning insmart cities, *Computer Communications*, *154*, 313–323.

Lee, J.H., Shin, J., & Realff, M.J. (2018). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers & Chemical Engineering*, *114*, 111-121.

Dowling, S., Schukat, M., & Barrett, E. (2018). Improving adaptive honeypot functionality with efficient reinforcement learning parameters for automated malware. *Journal of Cyber Security Technology*, *2*(2), 75-91.

Ray, S., & Turi, R.H. (1999). Determination of number of clusters in k-means clustering and application in colour image segmentation. *In Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, 137-143.

Schleich, M., Olteanu, D., & Ciucanu, R. (2016). Learning linear regression models over factorized joins. *In Proceedings of the International Conference on Management of Data*, 3-18.

Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications*, *4*(2), 13-19.

Spitzner, L. (2003). *Honeypots: tracking hackers* (Vol. 1). Reading: Addison-Wesley.

Wang, H.J., Guo, C., Simon, D.R., & Zugenmaier, A. (2004). Shield: Vulnerability-driven network filters for preventing known vulnerability exploits. *In Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communications*, 193-204.

Negi, P.S., Garg, A., & Lal, R. (2020). Intrusion detection and prevention using honeypot network for cloud security. *In 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* IEEE, 129-132.

Fraunholz, D., Pohl, F., & Schotten, H.D. (2017). Towards basic design principles for high-and medium-interaction honeypots. *In Proceedings of 16th European Conference on Cyber Warfare and Security*, 120.

Wang, H., & Wu, B. (2019). SDN-based hybrid honeypot for attack capture. *In IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, IEEE, 1602-1606.