Hybrid Model Based Intruder Detection System to Prevent Users from Cyber Attacks

Devendra Kumar Singh[1] and Manish Shrivastava[2]
devendra.singh170@gmail.com mannbsp@gmail.com
Dept of CSE, Central University, Bilaspur (CG) 495009, India

Summary
Presently, Online / Offline Users are facing cyber attacks every day. These cyber attacks affect user's performance, resources and various daily activities. Due to this critical situation, attention must be given to prevent such users through cyber attacks. The objective of this research paper is to improve the IDS systems by using machine learning approach to develop a hybrid model which controls the cyber attacks. This Hybrid model uses the available KDD 1999 intrusion detection dataset. In first step, Hybrid Model performs feature optimization by reducing the unimportant features of the dataset through decision tree, support vector machine, genetic algorithm, particle swarm optimization and principal component analysis techniques. In second step, Hybrid Model will find out the minimum number of features to point out accurate detection of cyber attacks. This hybrid model was developed by using machine learning algorithms like PSO, GA and ELM, which trained the system with available data to perform the predictions. The Hybrid Model had an accuracy of 99.94%, which states that it may be highly useful to prevent the users from cyber attacks.
Keywords-*IDS; Cyber Attacks; PSO; GA; ELM; KDD*

I. Introduction

Presently, users are spending most of their time to surf the internet for various purpose, people have extended their limits and resources by availing enormous facilities of internet. A cyber attack is kind of process which mainly triggers on computer systems, resources, networks or personal computer systems, by using different ways to copy, alter or change data or information processing. These cyber attacks are DoS, DDoS, MitM, Phishing, Drive by, Password, SQL injection, XSS, Eavesdropping, Birthday and Malware attacks [1]. These attacks create the final possibilities of being theft. The illegal access to any network with the intention to steal or to break the security of the network is termed as an Intrusion. The intrusion detection system is based on the assumption that the behavior of the intruder differs from particular user versus legitimate user in the ways of their activities [2] [3]. An IDS [4] is a device or software application which monitors a network for malicious activity or policy violations. Any detected activity or violation is typically reported either to an administrator or collected centrally by using a security information and event management (SIEM) system [5]. A SIEM system combines outputs from multiple sources and performs alarm filtering techniques to distinguish malicious activity from other activities [6]. Intrusion detection is a big challenging area in the domain of information security and very complicated and time-consuming process [7] [8]. Lots of IDS models are already developed but still the performance and accuracy of these models needs improvements [9].

The main objective of this research study is to improve the detection rate and accuracy of IDS model to find out the intruders in less time by proposing a predictive hybrid model to classify all types of security attacks [10] by using KDD 1999 dataset.

II. Existing work

Pradeep Singh et. al. 2018 [11] proposed a hybrid model for IDS. This hybrid model worked on Gaussian Mixture clustering with Random Forest Classifiers and k-Means clustering algorithm. Here, Hybrid model produced an accuracy of 99.7% when threshold value was 0.5. Mehrnaz Mazini, Babak Shirazi, Iraj Mahdavi et. al. 2018 [12] worked on A-NIDS using Artificial Bee Colony (ABC) and AdaBoost Algorithm using NSL-KDD 1999 along with ISCXIDS2012 dataset. AdaBoost Algorithm gives an accuracy of 98.90%. Mohan Allam, M. Nandhini, et. al. 2018 [13] worked on breast cancer using feature selection techniques by new binary teaching learning based optimization algorithm and they also proposed the FS-BTLBO algorithm and gave an accuracy of prediction which is 98.36%. Authors concluded that FS-BTLBO scheme is better than GA algorithm. Ruchi Makani, B.V.R. Reddy, et. al. 2018 [14] worked on MANAT's security from intruders through IDS. MANAT is dynamic

Manuscript received April 5, 2021
Manuscript revised April 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.4.33

allocation system. The machine learning-based IDS model for MANAT is developed by using ML-based Anomaly-based Detection Model. Erik Lejon et. al 2018[15] suggested to find the best method for anomaly detection in the press-hardening process. The author contributes through literature by selecting a suitable method for anomaly detection in the press-hardening process by evaluating three ML methods based on training time, prediction time, ease of implementation and easy performance. Sai Prasad Potharaju, M. Sreedevi, 2018 [16] worked on Existing filter-based feature selection techniques using data mining for IDS. The author suggested Novel M-clusters of feature selection and ranking framework. They worked on SONAR dataset (i.e. UCI Repository dataset) as well as tested 3, 4, and 5 clusters of features to obtain an accuracy of 33%, 25%, and 20% respectively. Anand Sukumar J V et. al. 2018 [17] compared between the two algorithms i.e. genetic algorithm and k-means algorithm respectively. They found that IGKA is better than a k-means algorithm and accuracy of k-means along with IGKA is 53.27% and 72.91% respectively. Chaouki Khammassi et. al. 2017 [18], In this paper authors proposed the model of GA-LR wrapper based upon decision tree classifiers technique and found theaccuracy of 81.42%. M.R. Gauthama Raman et. al. 2017[19], worked on features selection method by using adaptive and robust IDS techniques along with Hyper graph-based GA (HG-GA) and SVM classifiers. The above research reviews related to the use of IDS models based upon machine learning are very potential. This domain is very active and offers new solutions towards rapidly occurred challenges. The efforts in this area are vey less hence this research contribution may fill that gap.

III. Research methodology

A. Data Collection

The data used in this research paper derived from KDD 1999 dataset [20]. This data were compiled into CSV format containing 02 Lacks records along with 41 features of intruders with two class content and service features (table I). These features are mainly responsible for the transmission of dataset between the source and destination system. The authors used these features to find out the intruders [21] by proposed hybrid model. Every feature is responsible for sending the data in between the source and destination.

Table I. Kdd 1999 dataset features of intruders

Class	**Features of KDD'99**	**Class**	**Features of KDD'99**
Basic Features	Duration (F1)	Host Features	Count (F23)
	protocol_type (F2)		srv_count (F24)
	Service (F3)		serror_rate (F25)
	Flag (F4)		srv_serroe_rate (F26)
	src_bytes (F5)		rerror_rate (F27)
	dst_bytes (F6)		srv_rerror_rate (F28)
	Land (F7)		same_srv_rate (F29)
	wrong_fragt (F8)		diff_srv_rate (F30)
	urgent (F9)		srv_diff_h_rate (F31)
Content Features	Hot (F10)	Services Features	host_count (F32)
	num_fail_login (F11)		host_srv_count (F33)
	logged_in (F12)		h_same_sr_rate (F34)
	nu_comprom (F13)		h_diff_srv_rate (F35)
	root_shell (F14)		h_src_port_rate (F36)
	su_attempted (F15)		h_srv_d_h_rate (F37)
	num_root (F16)		h_serror_rate (F38)
	nu_file_creat (F17)		h_sr_serror_rate (F39)
	nu_shells (F18)		h_rerror_rate (F40)
	nu_access_files (F19)		h_sr_rerror_rate (F41)
	nu_out_cmd (F20)	--	--
	is_host_login (F21)	--	--
	is_guest_login (F22)	--	--

B. Data Selection and Preparation

From the collected dataset, a object oriented dataset was formed, by feature selection methods. The final selected dataset were processed, filtered, analyzed and prepared by MALAB [22]. The preprocessing of dataset was done Decision Tree which always provides high accuracy results.

C. Proposed Hybrid Model

Machine Learning is a very useful technique in which improvement comes through experiences along with the use of dataset. PSO, Genetic Algorithm and ELM are the various highly used variation of machine learning algorithms [23]. PSO is a feature reducing techniques used for feature reduction and improving the result of any machine learning model [24]. Genetic Algorithm is used for novel feature selection strategy which uses balanced and unbalanced data. ELM is a feed-forward neural network used for feature optimization. It is a single hidden layer network in which the hidden layer will be selected on random bas to optimize the features in minimum numbers [25]. ELM reduces the complexity in training by using the random values in wait and bias values. These values are fixed. Property of ELM is feature minimization by increasing the classification accuracy to improve the learning speed. In this work, increasing the efficiency of the learning process to the model and reducing upto 83% of the data space [26].
In this research paper, Authors proposed hybrid model [fig 1] which can describe and distinguish classed of data by using above mentioned machine learning algorithms. Here, the data set is divided into three groups (Table II). First group, the training data set equipped with 70% of the data, was taken into account to train the classifier. Second & third group, the testing & validation data sets equipped with 15% each of the data, which input into model for testing and validation based on the result of training phase. The proposed hybrid model predicted labels of classes along with the accuracy of classification.

Table II. Kdd1999 dataset divided into training, validation and testing dataset

Randomly Divided Dataset into 311029 Data Samples	Total Samples used in Model out of 311029 Data Samples
Group [I] Training Data 70%	217721
Group [II] Validation Data 15%	46654
Group [III] Testing Data 15%	46654

Fig 1 presents the proposed hybrid model based upon GA, PSO and ELM techniques. Dataset is provided to the combination of GA, ELM and PSO, ELM for reducing the number of features to detect intruders [26]. Further, the output of these two techniques input into proposed hybrid model to optimize the number of features.

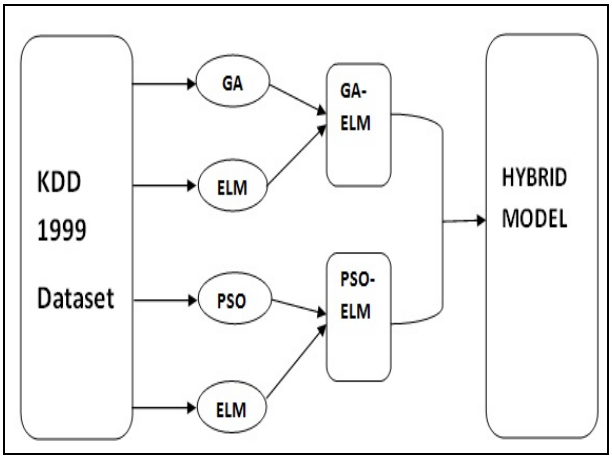

Fig 1. Proposed Hybrid Model

D. RESULTS AND EVALUATION

It is very important to control cyber attacks in the area of national defence, finance, insurance, private and public sectors. In this research study, a hybrid model based upon machine learning algorithms is proposed which searches intruders with minimum number of features by using KDD 1999 dataset in very less time. Once minimum number of features are sort out, it is easy for system/network administrator to control cyber attacks.

In this continuation, the confusion matrix is developed to describe the performance of proposed model on a particular set of test data for which true values are well known. Table III and figure 2 represents the confusion matrix of our system which is a essential part of our evaluation criteria. It is observed that, the total number of True Positive (TP) and True Negative (TN) predictions are high as the 97% of overall predictions. The cross validation results (Table IV) represents the classification report which includes accuracy, precesion, recall, F1 score and mean square error (MSE) [27] of our proposed hybrid model.

Table III. Confusion matrix

N=2089	Predicted "NO"	Predicted "YES"
Actual "NO"	350	8
Actual "YES"	38	1823

Table IV. Cross validation results

	Accuracy	Precision	Recall	F1 Score	MSE
0.0	0.97	0.9	0.97	0.95	0.022
0.1	0.97	1.0	0.97	0.95	

Fig 2. Confusion matrix contains true positive rates and false negative rates along with all classification of attacks

Fig 3 & 4, presents the performance comparison architecture of hybrid models tested in MATLAB 2018b with three parameters like number of features, accuracy and detention rate [28]. We observed that our proposed hybrid model performs very well in these performance metrics compared to existing models by minimizing the features to five (05) and achieving the higher accuracy rate (99.94%).

Finally, these five features (derived from table I) are **F2 (Protocol Type), F3 (Service), F5 (Source Bytes), F6 (Destination Bytes) and F23 (Count).** Now IDS systems can use these filtered features to quickly capture the information of intruders.

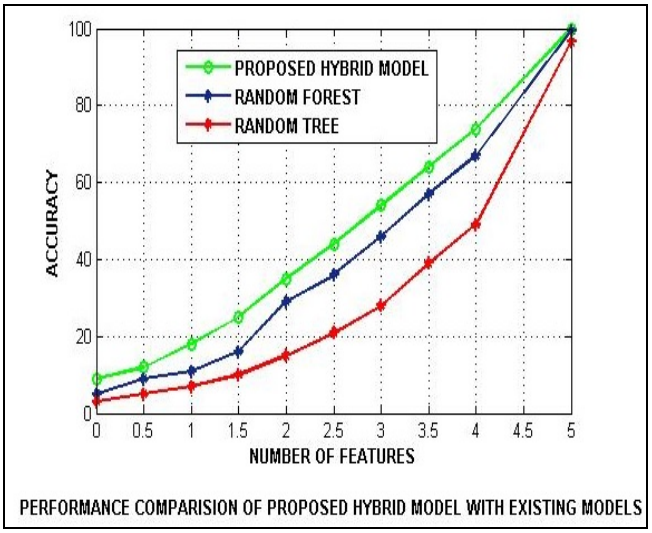

Fig 3. comparision of proposed model

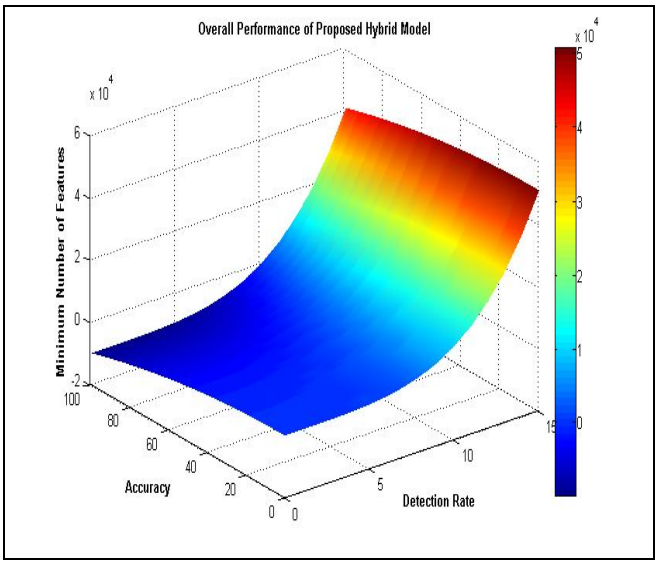

Fig 4. Overall performance of proposed model

V. Conclusion and future work

The prevention towards cyber attacks by using IDS is an emerging research area because now day's cyber attacks are very common to online/offline users on everyday basis. Attention must be given to prevent these attacks as they are affecting efficiency, resources and performing fraud activities. In this research paper, most of the cyber attacks are covered, i.e. Now, the intruders can be identified by using minimum number of features which can helps any IDS system to quickly spot the issue of cyber attacks. This model helps in this regard by

improving productivity and accuracy of available IDS systems.

In future, researchers may extend this hybrid model by in-cooperating other machine learning algorithms also like ANN and fuzzy theory to overcome the emerging issues like financial and privacy of smart phone users.

References

[1] Andreea Bendovschi: *Cyber-Attacks – Trends, Patterns and Security Countermeasures*, Procedia Economics and Finance. Science Direct, 28, 24-31 (2015)

[2] Nikolov, D., Kordev, I., & Stefanova, S.: *Concept for network intrusion detection system based on recurrent neural network classifier*. IEEE XXVII International Scientific Conference Electronics - ET, 1–4 (2018)

[3] Chary, K. C.: *Data Mining, Intrusion Detection System – A Study*. International Journal of Advanced Research in Computer Science, 3(1), 434–437 (2012)

[4] Khammassi Chaouki, Krichen Saoussen: *A GA-LR wrapper approach for feature selection in network intrusion detection*, In Computers and Security, Elsevier (2017)

[5] Viegas Felipe, Rocha Leonardo, Gonçalves Marcos, Mourão Fernando, Sá Giovanni Salles Thiago, Andrade Guilherme, Isac Sandin: *A Genetic Programming approach for feature selection in highly dimensional skewed data, In Neurocomputing*, Elsevier (2018)

[6] Gauthama Raman M. R., Somu Nivethitha, Kirthivasan Kannan, Liscano Ramiro, Shankar Sriram V. S.: *An efficient intrusion detection system based on hypergraph - Genetic algorithm for parameter optimization and feature selection in support vector machine, In Knowledge-Based Systems,* Elsevier (2017)

[7] Ariafar Elham, Kiani Rasoul: *Intrusion Detection System Using an Optimized Framework Based on Datamining Techniques*, In IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), Iran University of Science and Technology) – Tehran, Iran (2017)

[8] Saxena Aumreesh Ku, Sinha Sitesh, Shukla Piyush: *General study of intrusion detection system and survey of agent-based intrusion detection system*, In Proceeding - IEEE International Conference on Computing, Communication and Automation, ICCCA (2017)

[9] Jabbar, M. A., Aluvalu, R., & Reddy, S. S. S.: *Intrusion Detection System using Bayesian Network and Feature Subset Selection*. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 1–5 (2017)

[10] Mishra, V. P., & Shukla, B.: *Development of simulator for intrusion detection system to detect and alarm the DDoS attacks.* In 2017 International Conference on Infocom Technologies and Unmanned Systems: Trends and Future Directions, ICTUS, Institute of Electrical and Electronics Engineers Inc. Vol. 2018-January, pp. 803–806 (2018)

[11] Singh, P., & Venkatesan, M.: *Hybrid Approach for Intrusion Detection System; Hybrid Approach for Intrusion Detection System.* Proceeding of 2018 IEEE International Conference on Current Trends towards Converging Technologies (ICCTCT) (2018)

[12] Mazini, M., Shirazi, B., & Mahdavi, I.: *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. King Saud bin Abdulaziz University. (2018)

[13] Balikas, G., & Partalas, I.: *On the effectiveness of feature set augmentation using clusters of word embeddings*. In CEUR Workshop Proceedings Vol. 2226, 26–32 (2018)

[14] Allam, M., & Nandhini, M.: *Optimal feature selection using binary teaching learning-based optimization algorithm*. Journal of King Saud University - Computer and Information Sciences. King Saud bin Abdulaziz University (2018)

[15] Makani, R., & Reddy, B. V. R.: *Taxonomy of Machine Leaning Based Anomaly Detection and its suitability*. In Procedia Computer Science (Vol. 132, pp. 1842–1849). Elsevier B.V. (2018)

[16] Lejon, E., Kyösti, P., & Lindström, J.: *Machine learning for detection of anomalies in press-hardening: Selection of efficient methods.* In Procedia CIRP Vol. 72, pp. 1079–1083 (2018)

[17] Potharaju Sai Prasad, Sreedevi M.: *A Novel Subset Feature Selection Framework for Increasing the Classification Performance of SONAR Targets*, In Procedia Computer Science, Elsevier (2018)

[18] Sukumar, J. V. A., Pranav, I., Neetish, M., & Narayanan, J.: *Network Intrusion Detection Using Improved Genetic k-means Algorithm*. In 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI) 2441–2446 (2018)

[19] Borkar, A., Donode, A., & Kumari, A.: *A survey on Intrusion Detection System (IDS) and Internal Intrusion Detection and protection system (IIDPS).* In Proceedings of the International Conference on Inventive Computing and Informatics, ICICI 2017, Institute of Electrical and Electronics Engineers Inc. pp. 949–953 (2018)

[20] Knowledge discovery Dataset collected from UCI Repository http://kdd.ics.uci.edu/databases/kddcup99

[21] Hadri, A., Chougdali, K., & Touahni, R.: *Intrusion detection system using PCA and Fuzzy PCA techniques.* In International Conference on Advanced Communication Systems and Information Security, ACOSIS 2016 - Proceedings. Institute of Electrical and Electronics Engineers Inc. (2017)

[22] https://www.mathworks.com/products/matlab.html

[23] Mehmood, T.: Comparative Analysis Of Machine Learning Algorithms In Context Of Intrusion Detection. (2015)