# Efficient Host Based Intrusion Detection System Using Partial Decision Tree and Correlation Feature Selection Algorithm

F.Lydia Catherine [1], Ravi Pathak [2], V.Vaidehi [3]
[1] PG Student, Dept. of Information Technology, [2] Research Scholar, [3] Professor.
*Madras Institute of Technology, Anna University, Chennai*
lydiacatherinef@yahoo.com[1],pathak.ravi1989@gmail.com[2],vaidehi@mitindia.edu[3]

***Abstract*—System security has become significant issue in many organizations. The attacks like DoS, U2R, R2L and Probing etc., creating a serious threat to the appropriate operation of internet services as well as in host system. In recent years, intrusion detection system is designed to prevent the intruder in the host as well as in network systems. Existing host based intrusion detection systems detects the intrusion using complete feature set and it is not fast enough to detect the attacks. To overcome this problem, this paper proposes an efficient HIDS – Correlation based Partial Decision Tree Algorithm (CPDT). The proposed CPDT combines Correlation feature selection for selecting features and Partial Decision Tree (PART) for classifying the normal and the abnormal packets. The algorithm is implemented and has been validated within KDD'99 dataset and found to give better results than the existing algorithms. The proposed CPDT model provides the accuracy of 99.9458%.**

***Keywords—IDS, Intruder, R2L, CFS, Probing, DoS, U2R.***

## I. INTRODUCTION

In recent years, there are lots of attacks on the computer system. As the internet services have drastically grown, there are more threats and malicious intruders to compromise the parameters such as confidentiality, security, and availability of host system.Intrusion Detection System(IDS) is a method to keep the host system in secure condition.The IDS captures and analyzes the network or host system events for abnormal activities or unauthorized access and produces alerts to the user. Intrusion Prevention System (IPS) is a security appliance that can monitor the system behavior for malicious activity and attempt to stop an intruder. Intrusion can take place in host systems and in network systems. Host-based IDS (HIDS) is designed to protect the host system from unauthorized access and prevents destruction of the user's information. It includes manipulation of system calls, modification of file systems, privilege escalation, unauthorized logging into the system and unauthorized access to the highly confidential files and malware which can affect the state of the system. Network-based IDS (NIDS) provides protection from the attacks by analyzing the network traffic.Attacks are classified into four types [1] - Denial of Service (DoS), Remote to Local (R2L), User to Root (U2R) and Probing. DoS attack is used to make a host system or the network unavailable to its users. R2L is used to perform the unauthorized access to victim machine by the attacker. In U2R, attacker has local access to the victim machine and attempt to gain super user privileges. Probing is used to scan the ports and collects the information about the system activity. To avoid these vulnerabilities IDS is implemented in the host system.

The IDS provides one of the most promising paths towards network robustness. IDS are used to detect the several types of malicious activities that can violate the rules and trust of the host systems. IDS can further classify as Anomaly based intrusion detection and Signature based intrusion detection. Anomaly-based IDS determines normal network behavior like bandwidth range, types of protocols, ports and a device used to connect each other and alertswill be sent to the network administrator or user, when inconsistent traffic is detected. On the other hand, Signature-based IDS monitors packets in the network and compares them with pre-configured and pre-identified attack behaviors, called signatures.

The dataset created by combining the information in the system logs. The major challenges in the existing KDD'99 based IDS arethat data set is of large number of attributes which increases the processing time of system.The accuracy of the existing classifiers like C4.5, Random forest etc., is also low. The KDD'99 dataset consistsof 42 features [6]. So, it requires more time to process and it takes longer time to detect the known attacks. To overcome this issue CPDT intrusion detection system is proposed in this paper. The proposed CPDT based IDS uses Correlation Feature Selection (CFS) to extract the features from the KDD'99 dataset and the dataset with the extracted features is given to the Partial decision tree (PART) algorithm which classifies the normal and abnormal behaviour of the host system.

This paper is organized into five sections. Section II deals detailed study of the literature review. Section III deals with the proposed work that includes 3 stages namely dataset collection, recognizing the attack scenario in that dataset anddetection of known attacks using CPDT. Section IV deals with the results and discussions. Section V gives the conclusion for the paper.

## II. RELATED WORK

This section deals the related works on HIDS and the existing classification algorithms applied on KDD'99 dataset. Enormous amount of work has been conducted on KDD 99

978-1-4799-4989-2/14/$31.00 

Intrusion Detection datasets for classification of different attacks. Tavallaee [6]worked with several clustering algorithms for anomaly detection using KDD'99 data set. It has tried to identify the unknown attack however the model could achieve maximum accuracy of 80%. Ding et.al., [14] introduced a new technique that combines misuse detection system with anomaly detection system. Their proposed system contains 3 sub modules namely misused detection module, anomaly detection module and signature generation module. Misused detection module is used to detect the known attacks by using snort. Anomaly detection module detects the unknown attacks and signature generation module extracts signature of the attacks that are detected by Anomaly Detection module and maps the signature into snort. Accuracy of system was 87% at cost of higher time complexity.

The IDS designed by Massimo Ficcoet.al., [8] collects information at several architectural levels and performs complex event correlation based on complex event processing engine. This intrusion detection system implements a comprehensive alert correlation workflow for detection and diagnosis of complex intrusion scenarios in large scale complex infrastructures. Their model could attain an accuracy of 91%.

The intrusion model proposed byYing et.al., [5] enhances the detection of intrusion by combining two detection techniques namely log file analysis technique and Back Propagation neural network model technique. However they could achieve accuracy of 93% at cost of longer detection time.

Log attribute selected module is proposed by Wang et.al., [16]. In this C4.5 algorithm, Naïve Bayes, Random tree and Random forest algorithm are used to analyze the features in the training samples. The Classification algorithm C4.5 creates the decision tree using the information in the KDD'99 dataset. It constructs the decision tree with the minimum number of nodes. The C4.5 produces the better results than the other three algorithms with an accuracy of 98%.

Leu,De-Zhan et.al,[7] applied association and sequential rules mining for detecting the intrusion. However they can achieve 85% accuracy with the proposed model.RadhikaGoelet.al[13] proposed a novel hybrid model for anomaly detection and misuse detection. C4.5 decision tree used for misuse and CBA classifier is applied for anomaly detection. C4.5 splits the packet into normal and abnormal. The normal packet is sent to theanomaly detector and abnormal packets are sent to a tree based classifier for identifying the type of the attacks. But few deficiencies of KDD'99 dataset cannot be solved completely.It achieves an accuracy of 98.5%.

Mrudulagudadeet.al., [10]proposed new ensemble decision tree for IDS. It reduces the computational cost than theexisting methods which achieves 93% accuracy. K.Nageswararao et.al.[9] applied CART algorithm on KDD'99 dataset to analyze the attacks behavior. CART algorithm achieves the accuracy of 94%. Mansour et.al., [7] proposed an intrusion detection system with the unsupervisedneural network called as Kohonen maps. The performance based ranking method was used to reduce the false positive rates. Only one record of the dataset is removed from KDD'99 dataset. The results are compared before and after removal of the dataset. Kohonen map could provide an accuracy of 88%.

Nilimapatilet.al., [11] proposed an IDS and two algorithms are used namely C5.0 and CART. From the statistical analysis it is proved that C5.0 provides the more accuracy when compared to the CART. C5.0 attained an accuracy of 98.5%. The limitations of the existing work are the less accuracy, requires more computational time and more memory space to store the 42 features of a dataset. These limitations are solved in the proposed CPDT system.

## III .PROPOSED CPDT BASED IDS

This section contains the architecture of CPDT the proposed intrusion detection system followed by the PART algorithm. The proposed CPDT based IDS has three stages namely dataset collection, recognition of the attack scenario and detection of known attacks.The proposed CPDT based IDS is shown in the Figure 1. The functionality of each module is explained below:

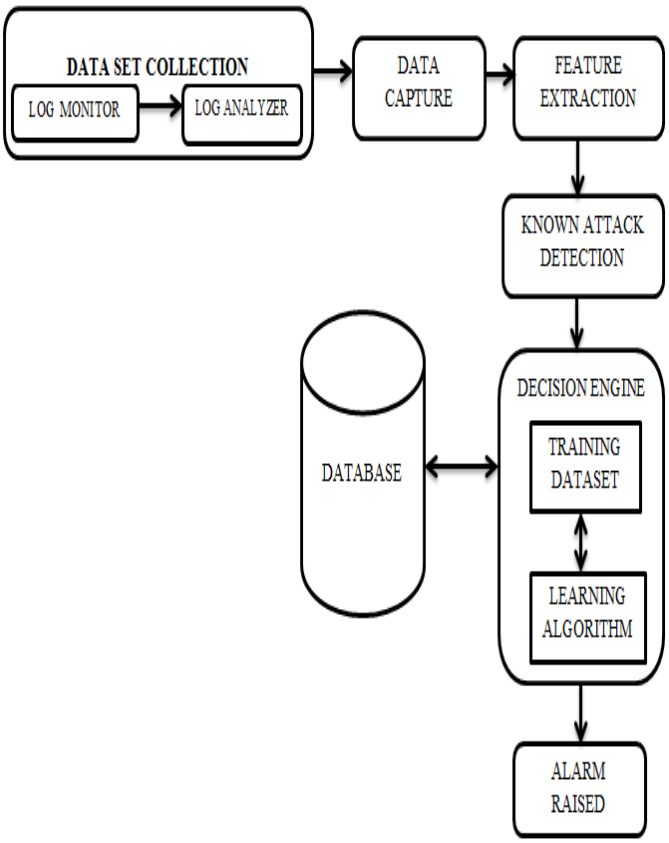


Fig. 1. Design of Proposed CPDT System

### *A. DATASET COLLECTION:*

In this stage data packets are collectedby monitoring the system logs. The collected information is maintained in the database.

*i) Log Monitor:*Log Monitor monitors the events of the host system.

*ii) Log Analyzer:* Log Analyzer module analyzes the behavior of the system.

### *B. RECOGNITION THE ATTACK SCENARIO*:

During an attack scenario, it is more likely that the attacker first performs some knowledge gathering steps which consists of commands that enables to acquire the knowledge about the target system.

*C. DETECTION OF KNOWN ATTACKS:*

Before classifying the packets into normal and abnormal using PART algorithm, the features should be extracted using the CFS technique.

*i) Feature Extraction:*

Feature Extraction is used to reduce the inputs to a reasonable size for analyzing and processing the inputs. It is most important part of IDS. Features of the KDD'99 dataset should be extracted before testing. The CFS is used as the feature extraction algorithm. It is based on the hypothesis "Good feature set contains features that are highly correlated with classification but not correlated with each other".The equation (1) gives the merit ($M_k$) of a feature subset consisting of k features.

$$M_k = \frac{k.t_f}{\sqrt{k+(k-1)r_{ff}}} \qquad (1)$$

Where, $t_f$ is the average value of feature classification correlations in KDD'99 dataset and $r_{ff}$ is the average value of all feature correlations in the KDD'99 dataset. The equation (2) defines the CFS criterion ($CFS_{fs}$).

$$CFS_{fs}=\max \frac{(t_{f1}+t_{f2}+\ldots+t_{fk})}{(\sqrt{k+2(r_{f1f2}+\ldots+r_{fifj})}} \qquad (2)$$

Where, the $t_{fi}$and$r_{fifj}$are the correlations. Some of the features extracted are the protocol type, service, src_bytes, dest_bytes,land,attack,wrong_fragment,root_shell,count,diff_srv_rate,dest_host_same_src_port_rate.

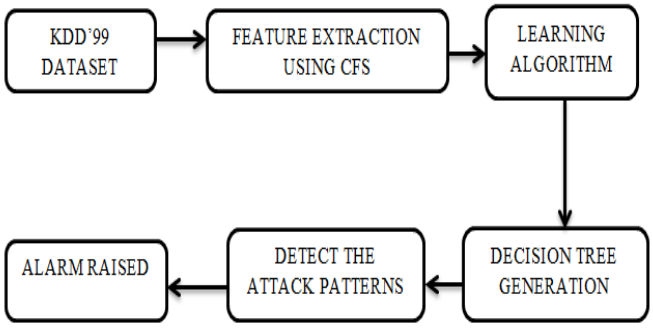


Fig. 2. Detecting the attack patterns from the dataset.

*ii) PART algorithm:*

The PART is the combination of both C4.5 and RIPPER classification algorithm. The CFS with PART algorithm steps is given Algorithm 1.Figure 2 presents the block diagram for detecting the attack patterns in the dataset.The KDD'99 dataset contains 42 features and those features are extracted using CFS.Finally 12 features are selected by using the CFS. The packets with those 12 features are given to the PART algorithm. PART divides the dataset into subsets and the subsets are expanded further till the leaf node is reached. The nodes are replaced by the leaf nodes.

**Algorithm 1: CFS-PART**

**Input**

I. . . set of training instances

**Output**

RedF. . . a reduced subset of features

1. Start
2. Get the dataset with features
3. Initialise the CFSfs(I) with the training instances
4. RedF==0
5. Set Featureset=CFS(I)
6. For each feature in feature set, repeat the following steps
   i)Extract the features of feature set and store it in the EXFEATURE
   ii) Merge the EXFEATURE with the RedF.
7. End.
8. Return RedF.
9. Split the datasets into subsets with RedF
10. Expand the first subset.
11. While {subsets that have not been expanded and all the subsets expanded are reached the leaf level}
12. Choose next subset to be expanded.
13. Expand the second subset.
14. Repeat the expansion process till all the subsets are expanded
15. Replace node by a leaf.
16. Stop.

## IV. EXPERIMENTATION AND EVALUATION

This section explains about the experimentation of proposed CPDT based IDS.Snort is a rule based engine and it is used to capture the packets.The KDD'99 dataset is used to validate the effectiveness and accuracy of proposed CPDT system. The KDD'99 intrusion detection datasets are based on KDD'98 initiative which is used as a benchmark for evaluating the different IDS.The proposed CPDT based IDS system is validated in a network consisting of several target machines running various operating systems and services.Another three machines are used to perform spoofing to generate traffic. Finally, thereis a sniffer that records all the traffic using the TCP dump format[6].The total simulated period is seven weeks. Using CFS 42 features of KDD'99 dataset is extracted. The extracted features with its own characteristics are listed in the Table I. These features are divided into two greater types that are continuous and discrete. There are 22 types of attacks have been focused in KDD'99 dataset. The attacks are back dos, buffer_overflow u2r, ftp_write r2l, guess_passwd r2l, imap r2l ipsweep probe, land dos, load module u2r, multihop r2l, neptune dos, nmap probe, perl u2r, phf r2l, pod dos, portsweep, probe, rootkit u2r, satan probe, smurf dos, spy r2l, teardrop dos, warezclisent r2l, warezmaster r2l. The KDD'99 database contains the set of data to be audited which includes

a wide variety of intrusions in the host system. Table II shows the frequency of occurrence of popular attacks.

**TABLE IEXTRACTED FEATURES WITH ITS BEHAVIOR**

| Feature Name | Description | Type |
|---|---|---|
| Protocol_Type | Type of the protocol e.g. udp, http etc. | Discrete |
| Service | Network service on the destination | Discrete |
| Src_bytes | Number of data bytes from source to destination | Continuous |
| dest_bytes | Number of data bytes from destination to source | Continuous |
| Flag | Normal or error status of the connection | Discrete |
| Land | 1 if connection is from/to the same port; 0 otherwise | Discrete |
| Count | Number of connections to the same host as the current connection | Continuous |
| Diff_srv_rate | Percentage of connections to different services | Continuous |

**TABLE II CHARACTERISTICS OF KDD'99 INTRUSION DETECTION DATASET**

| Dataset | DoS | U2R | R2L | Normal |
|---|---|---|---|---|
| 10% KDD'99 | 391458 | 52 | 1126 | 97277 |
| Corrected KDD | 229853 | 70 | 16347 | 60593 |
| Whole KDD dataset | 3883370 | 52 | 1126 | 972780 |

*A.WEKA:*

WEKA is a collection of algorithms for data mining tasks [9]. It contains the tools for data pre-processing, classification, regression, clustering, association rules, and visualization. The features that are extracted using the CFS areProtocol_type, Service, Flag,Src_bytes, Dest_bytes, Land, Wrong_fragment, Root_shell,Count,Diff_srv_rate,attack,Dest_host_same_src_port_rate. The total number of instances that are taken for testing is 494021. The efficiency of the proposed CPDT based IDS is 99.9458%. So, it is found that the results are more accurate than the existing systems [2]. The incorrectly identified instances are 0.05%. The efficiency of CPDT based IDS is shown in Figure 3.

| | | |
|---|---|---|
| Correctly Classified Instances | 493753 | **99.9458%** |
| Incorrectly Classified Instances | 268 | **0.0542%** |
| Kappa Statistic | 0.9991% | |
| Mean absolute error | 0.0001% | |
| Root mean squared error | 0.0065% | |
| Relative absolute error | 0.1418% | |
| Root relative squared error | 4.0808% | |
| Coverage of cases (0.95 level) | 99.9759% | |
| Mean rel. region size (0.95 level) | 4.3624% | |
| Total Number of Instances | 494021 | |

Fig. 3. Efficiency produced by proposed CPDT

The accuracy of the proposed CPDT based IDS is given for different classesin Table III. Confusion matrix is a table that allows visualizing the performance of the IDS. The confusion matrix generated by the proposed CPDT based IDS is shown in Table IV. In the confusion matrix,a denotes the normal class, b denotes the buffer overflow, c denotes the load Module, d denotes the perl, e denotes the Neptune, f denotes the rootkit, and f denotes the warezclient. The equations for performance measures in terms of precision, recall, False Positive Rate(FPR), and the Detection Rate are shown in equations 3, 4, 5 and 6 respectively[13].

$$\text{Precision} = \frac{(TP)}{(TP+FP)} \quad (3)$$

$$\text{Recall} = \frac{(TP)}{(TP+FN)} \quad (4)$$

$$\text{FPR} = \frac{(FP)}{(FP+TN)} \quad (5)$$

$$\text{Detection Rate} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (6)$$

Receiver Operating Characteristic (ROC) curve shows the efficiency of the detection rate of the CPDT based IDS. ROC curve is plot of the true positive rate (TPR) vs. false positive rate (FPR). The performance of the system is validated with the precision, recall and the detection rate. Precision is the fraction of the attacks that are expected and recall is the fraction of the relevant instances that are obtained. The ROC curve is plot by taking FPR in X-axis and TPR in Y-axis.The ROC curve for normal packets is shown in Figure 4. As the ROC curve is closer to the left hand border and then the top border it is inferred that the results are more accurate. The ROC curve for buffer overflow is presented in Figure 5. The area under ROC for buffer_overflowis 0.9 and it indicates that the detection of the smurf is accurate.Figure 6 shows the efficiency of the proposed CPDT based IDS. The proposed CPDT based IDS is compared with the C4.5, CART, 0R, 1R. The proposed CPDT based IDS provides the accuracy of 99.9458% than the existing IDS. The ROC curve is plotted by taking false positive rate in the X-axis and the True positive rate in the Y-axis. The ROC curve (Figure 7) of the proposed CPDT based IDS is closer to the top border and it is inferred that the detection rate is high.

**TABLE IIIACCURACY OF PROPOSED CDPT BASED IDS FOR DIFFERENTCLASSES**

| TRUE POSITIVE | FALSE POSITIVE | PRECISION | RECALL | F-MEASURE | MCC | ROC AREA | PRC AREA | CLASS |
|---|---|---|---|---|---|---|---|---|
| 1.000 | 0.000 | 0.999 | 1.000 | 0.999 | 0.999 | 1.000 | 1.000 | Normal |
| 0.700 | 0.000 | 0.778 | 0.700 | 0.737 | 0.738 | 0.900 | 0.666 | Buffer_overflow |
| 0.444 | 0.000 | 0.308 | 0.444 | 0.364 | 0.370 | 0.889 | 0.226 | Load module |
| 0.667 | 0.000 | 0.500 | 0.667 | 0.571 | 0.577 | 0.833 | 0.667 | Perl |
| 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | Neptune |
| 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.700 | 0.066 | rootkit |
| 0.987 | 0.000 | 0.990 | 0.987 | 0.989 | 0.989 | 0.997 | 0.986 | Warezclient |

**TABLE IV CONFUSION MATRIX**

| a | b | c | d | e | f | g | Representation of Class |
|---|---|---|---|---|---|---|---|
| 97236 | 1 | 4 | 1 | 8 | 1 | 3 | a→Normal Packet. |
| 5 | 21 | 2 | 0 | 0 | 0 | 0 | b→Buffer Overflow. |
| 1 | 0 | 0 | 2 | 0 | 0 | 0 | c→Load Module. |
| 3 | 0 | 0 | 0 | 107196 | 0 | 0 | d→Perl. |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | e→Neptune. |
| 5 | 2 | 0 | 4 | 0 | 0 | 0 | f→Rootkit. |
| 12 | 0 | 0 | 0 | 0 | 0 | 0 | g→Warezclient. |

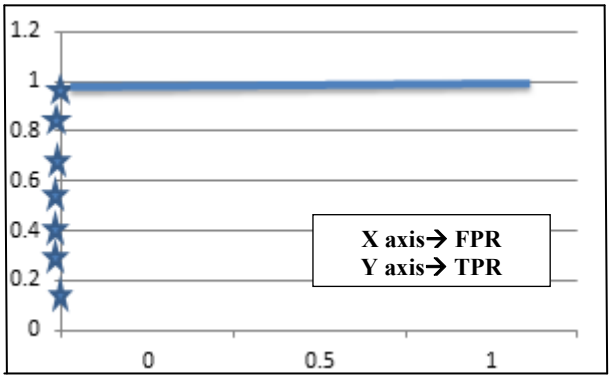


Fig. 4. ROC graph for normal packets

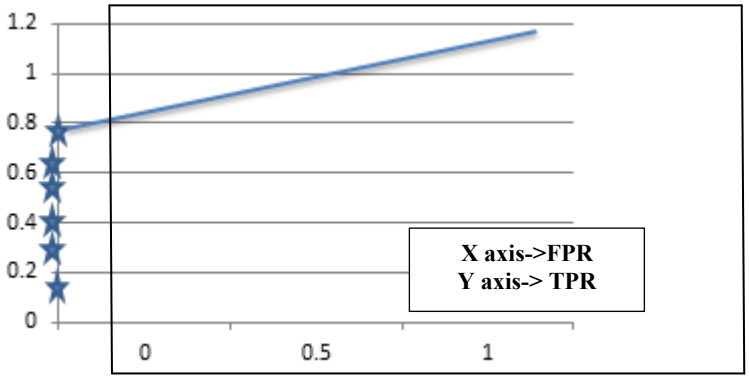


Fig. 5. ROC graph forbuffer_overflow

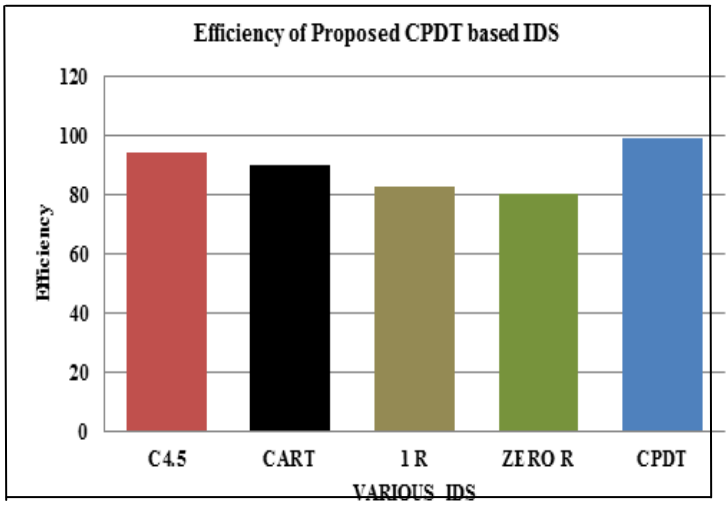


Fig. 6 Efficiency graph

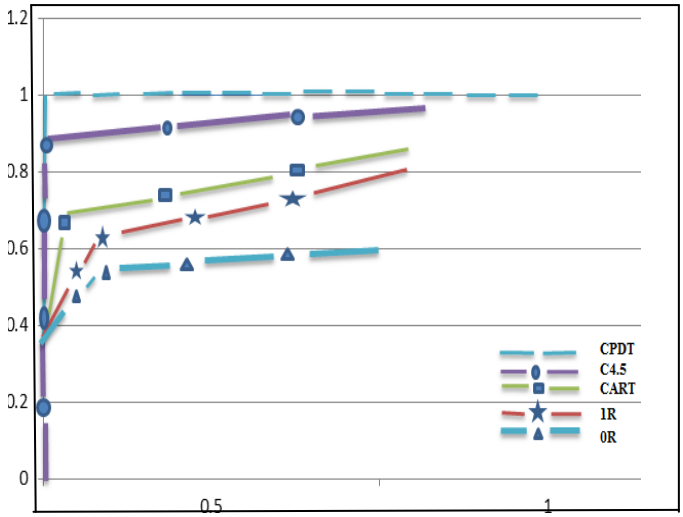


Fig. 7 ROC curve

## V. CONCLUSION

The CPDT based IDS is proposed in this paper. It is implemented in JAVA.The proposed system handles the attacks such as DoS, U2R, R2L and Probing. The proposed CPDT based IDS combines CFS for selecting the features and PART for classifying the normal and the abnormal packets in the KDD'99 dataset. The proposed system uses only less number of features for detecting intrusion.So the processing time is less.From the experimental results, it is proven that the proposed system is found to have better performance in terms of detection accuracy and the detection time and better for the real time environment. In future, this work will be integrated with the Complex Event Processing (CEP) engine to detect the unknown attacks.

## REFERENCES


[1] Amit Kumar, Harish Chandra Maurya, Rahul Misra(2013, April). Research Paper on Hybrid Intrusion Detection System. In *International Journal of Engineering and Advanced Technology*.

[2] Bin Hamid Ali, F. A., & Len, Y. Y. (2011, September). Development of host based intrusion detection system for log files. In *Business, Engineering and Industrial Applications (ISBEIA), 2011 IEEE Symposium on* (pp. 281-285). IEEE.

[3] Hoang, X. D., Hu, J., &Bertok, P. (2009). A program-based anomaly intrusion detection scheme using multiple detection engines and fuzzy inference. *Journal of Network and Computer Applications*, *32*(6), 1219-1228.

[4] Zhao, L., Fang, X., & Dai, Y. (2009, July). A novel adaptive intrusion detection approach based on comparison of neural networks and idiotypic networks. In *Nonlinear Dynamics and Synchronization, 2009. INDS'09. 2nd International Workshop on* (pp. 203-208). IEEE

[5] Ying, L., Yan, Z., & Yang-jia, O. (2010, April). The design and implementation of host-based intrusion detection system. In *Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on* (pp. 595-598). IEEE.

[6] Tavallaee, M., Bagheri, E., Lu, W., &Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In *Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009*.

[7] Chen, Y., Hwang, K., & Ku, W. S. (2007). Collaborative detection of DDoS attacks over multiple network domains. *Parallel and Distributed Systems, IEEE Transactions on*, *18*(12), 1649-1662.

[8] Ficco, M., & Romano, L. (2011, June). A generic intrusion detection and diagnoser system based on complex event processing. In *Data Compression, Communications and Processing (CCP), 2011 First International Conference on* (pp. 275-284). IEEE.

[9] Hmida, M. B. H., &Slimani, Y. (2010). Meta-learning in grid-based data mining systems. *International journal of communication networks and distributed systems*, *5*(3), 214-228.

[10] Gudadhe, M, Prasad, P., &Wankhade, K. (2010, September). A new data mining based network intrusion detection model. In *Computer and Communication Technology (ICCCT), 2010 International Conference on* (pp. 731-735). IEEE.

[11] K.NageswaraRao, D.Rajya Lakshmi, T.VenkateswaraRao (2012, March). Robust Statistical Outlier based Feature Selection Technique for Network Intrusion Detection" *International Journal of Soft Computing and Engineering, 2231-2307.*

[12] Patil, N., Lathi, R., &Chitre, V. (2012, June). Comparison of C5. 0 & CART classification algorithms using pruning technique. In *International Journal of Engineering Research and Technology* (Vol. 1, No. 4 (June-2012)). ESRSA Publications.

[13] PayamEmamiKhoonsari and Ahmad Reza Motie. (2012, Oct). Comparison of efficiency and robustness of ID3 and C4.5 algorithms using dynamic test and training datasets. In International *journal of machine learning and computing*.

[14] Goel, R., Sardana, A., & Joshi, R. C. (2012). Parallel Misuse and Anomaly Detection Model. *IJ Network Security*, *14*(4), 211-222.

[15] Ding, Y. X., Xiao, M., & Liu, A. W. (2009, July). Research and implementation on snort-based hybrid intrusion detection system. In *Machine Learning and Cybernetics, 2009 International Conference on* (Vol. 3, pp. 1414-1418). IEEE.

[16] YongengWang,Xiaoming Zhang. (2012, May). Complex Event processing over probabilistic event streams. In 9th*International Conference on Fuzzy Systems and Knowledge Discovery* (pp.1489-1493).IEEE.