INTERNATIONAL RESEARCH JOURNAL OF
HUMANITIES AND INTERDISCIPLINARY STUDIES
(Peer-reviewed, Refereed, Indexed & Open Access Journal)
DOI : 03.2021-11278686 ISSN : 2582-8568 IMPACT FACTOR : 5.71 (SJIF 2021)

Intrusion Detection Using Information Gain Feature Selection and Classification

[1]Senthilnayaki B, [1]G. Mahalakshmi, [1]J. Duraimurugan, [1]N. Anbarasi, [2]S. Prasath Kumar
[1]Department of Information Science and Technology, Anna University, Chennai (India)
[2]Department of ECE, Sri Sam Ram Institute of Technology, Chennai (India)
nayakiphd@auist.net

DOI No. 03.2021-11278686 **DOI Link ::** https://doi-ds.org/doilink/11.2021-73936339/IRJHIS2111023

Abstract:

With the expansion of the Internet, the number of assaults has skyrocketed, and Intrusion Detection Systems (IDS) have emerged as a critical component of information security. The goal of an intrusion detection system (IDS) is to assist computer systems in dealing with assaults. This anomaly detection system builds a database of typical behavior and deviations from it, which it uses to trigger when intrusions occur. IDS model is divided into two types based on the data source: host-based IDS and network-based IDS. Individual packets passing over the network are monitored in network based IDS, whereas actions on a particular computer or host are studied in host based IDS. The feature selection aids in the reduction of categorization time. It has been suggested and implemented in this work to successfully detect assaults. A novel feature selection method based on Information Gain Ratio dubbed Optimal Feature Selection has been developed and implemented for this purpose. The KDD Cup dataset is used to select the best number of features using this feature selection technique. In addition, the data set was effectively classified using two classification techniques: Support Vector Machine and Rule Based Classification. This method is highly effective at identifying DoS assaults and lowering the number of false alarms. The suggested feature selection and classification methods let the IDS identify assaults more effectively.

Keywords: *Intrusion Detection, Information Gain, Feature Selection Technique, Classification.*

1. Introduction:

Security has been a major problem in many sectors since computers have been networked together with a huge user base. Security for networks has become essential due to the fast expansion of internet communication and the availability of tools to infiltrate the network. The data contained in databases is not adequately protected by current security procedures. Many additional technologies, including as firewalls, encryption, and authorization systems, can provide security, but they are still vulnerable to hacker assaults exploiting system vulnerabilities.To defend these systems from hackers, a novel Intrusion Detection System has been designed and built in this project work, which uses the KDD cup data set to identify assaults using a basic feature selection method and SVM approach. The extraction of hidden predictive information from huge databases is known as data mining. It's a promising new technology that allows organizations to concentrate on the most relevant data in their data warehouses. Any type of information repository can benefit from data mining. When applied to different types of data, however, methods and methodologies may differ.

The internet has recently become a part of everyday life.Current internet-based information processing systems are vulnerable to a variety of attacks, resulting in a variety of damages and considerable losses. As a result, the value of information security is rapidly increasing. The most fundamental aim of network security is to create defensive networking systems that are protected against unwanted access, usage, disclosure, interruption, modification, or destruction. Furthermore, network security reduces the risks associated with key security objectives such as confidentiality, integrity, and availability.

2. Related Work:

With the introduction of the internet, network security has been a hot topic in recent years. Many publications exist in the literature that discuss Intrusion Detection Systems[1][2][3]. Intruder detection systems (IDSs) are used to identify intruder assaults. Sindhu et al [1] presented a genetic-based feature selection approach to reduce the classifier's computing complexity. Lee et al [2] developed an adaptive data mining technique for intrusion detection based on association criteria and frequent events obtained from audit data. Using a multiple-level hybrid classifier, Xiang and Lim [3] presented a misuse IDS.

Senthilnayaki et al [4] introduced an intrusion detection system (IDS) that employs feature selection and classification based on information acquisition. One of their approach's drawbacks is that it lengthens the training period. Sarasamma et al. [5] introduced a new multilevel hierarchical Kohonen network for detecting network intrusions. To train and test the classifier, they randomly chose data points from KDD Cup 99. Jianping Li et al [6] presented a new approach for choosing relevant feature sets for network intrusion detection based on the Continuous Random Function (CRF).

In the literature for IDS, there are several classification methods based on SVM. For example, Snehal A. Mulay et al [7] presented a Tree Structured Multiclass SVM method for effectively categorising data. Preprocessing is discussed in a number of publications in the literature [8][9][10]. Instead than calculating them exactly at the cost of decreased performance, time and space complexity, most real-world issues require an optimal and acceptable solution. The feature selection search began with a null set of features, which were added one by one, or with a complete set of features, which were removed one by one. In order to create an IDS, Li et al [12] presented a wrapper-based feature selection method.

The statistical approach for evaluating the large KDD Cup dataset is addressed by the feature selection algorithm developed by senthilnayaki et al [11]. Many books and articles exist in the literature that cover categorization strategies and tools [1][17][18][19]. Support Vector Machines (SVM) are classifiers that were created with binary classification in mind. For IDS, Debar et al [13] created a Neural Network (NN) model.By introducing membership to each data item, Du Hongle et al [14] suggested an enhanced v-FSVM. Senthilnayaki et al. [15] proposed a new learning approach for network intrusion detection based on a modified J48 classifier and ID3 algorithm, which identifies effective attributes from the training dataset, calculates conditional probabilities for the best attribute values, and correctly classifies all the examples in the training and testing dataset. [16] Suggested an intrusion detection system based on SVMs. Mahalakshmi G et al.[21] proposed a system using Convolutional Neural Network (CNN) to detect the intrusion in the network. The UNSW-NB15 dataset is used for the detection process. The system is capable of identifying the few attacks on the network.

3. Proposed System Implementation and Results:

The records from the KDD'99 cup data set are collected by the data collecting agent. This information is passed to the data preparation module, where it is preprocessed. The data acquired from the cup dataset might be either normal or attacked data [20].Preprocessing techniques are required for data reduction since processing large amounts of network traffic data with all attributes essential to detect attackers in real time and give preventive strategies is rather difficult.The application of rules fired utilizing the rule system triggered by the intelligent agents improves judgments on anomalous intrusion detection and prevention in this system.The major benefit of utilizing rules with a knowledge base is that it aids in making effective intrusion decisions.The SVM is a learning machine capable of binary classification and regression estimation. Because of two essential features, they are becoming increasingly attractive as a new paradigm of categorization and learning. To begin, unlike other classification approaches, SVM focuses on minimising anticipated error rather than classification error. Second, SVM uses mathematical programming's duality theory to create a dual issue that can be solved using efficient computing approaches.

The Information Gain Ratio was used to build this method for attribute selection. The data set D is split into n number of classes Ci to do this. The agent selects the characteristics Fi with the greatest number of non-zero values, and the Information Gain Ratio (IGR) is calculated using the following equations:

$$\text{Info (D)} = -\sum_{j=1}^{m} \left[\frac{freq(C_j,D)}{|\,D\,|}\right] log_2\left[\frac{freq(C_j,D)}{|D|}\right](1)$$

$$\text{Info (F)} = \sum_{i=1}^{n} \left[\frac{|F_i|}{|F|}\right] * info\ (F_i)(2)$$

$$\text{IGR } (\text{A}_\text{i}) = \left[\frac{Info(D) - Info\ (F)}{Info(D) + Info\ (F)}\right] * \ 100 \quad (3)$$

The following are the steps of the optimum feature selection algorithm.

Algorithm: Attribute Selection Algorithm Based on Intelligent Agents

Set of 41 characteristics from the KDD'99 Cup data set as input

R (reduced collection of characteristics) as an output

The algorithm's steps are as follows:

Step 1: Choose the characteristics that have a range of values.

Step 2: Using equation 1, calculate the Info(D) values for the specified characteristics.

Step 3: Choose the characteristics that have the most non-zero values.

Step 4: Using equation 2, calculate the Info(F) value for the characteristics chosen in step 3.

Step 5: Using equation 3, calculate the IGR value.

Step 6: Choose the characteristics based on the IGR value

The OFS algorithm has chosen ten key properties for detecting assaults effectively and reducing calculation time.

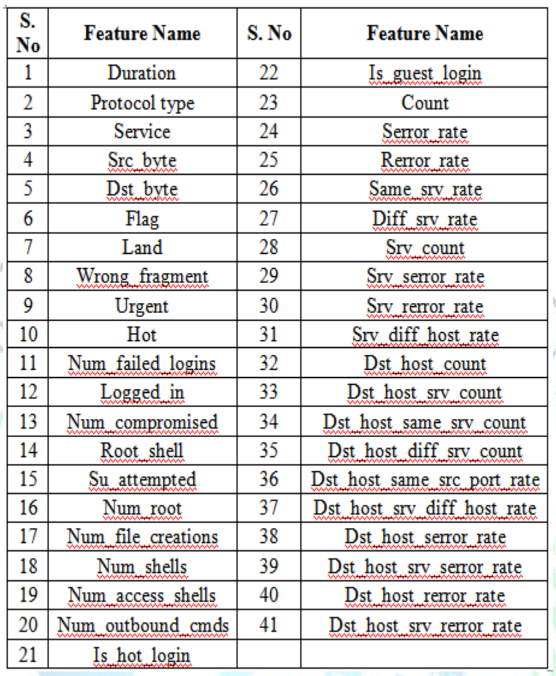
The pseudo code for selecting the best features is given below.

```
Input the data set
for each column in the data set
    Select non-varying columns
endfor
for each non-varying columns
    calculate frequency of each value in the data set
    calculate info(d)
endfor
for each column with maximum no. of non-zero values
    calculate frequency of each value
calculate info(f)
endfor
for each column
calculate IGR value
endfor
```

The OFS implementation algorithm is described in this section.

Optimal Feature Selection: Calculating IGR values with standard feature selection methods takes a long time. As a result, in this paper, a novel feature selection method called Optimal Feature Selection is suggested and implemented, which decreases computing time. This approach computes the Information Gain Ratio (IGR) value for the data set's various characteristics. It reduces the number of columns based on the IGR value. OFS improves detection accuracy while lowering false alarm rates.The simulated assaults are divided into four categories: Denial of Service (DoS), User to Root (U2R), Remote to Local (R2L), and Probe attacks.

Table1. The 41 features in KDD'99 dataset



S. No	Feature Name	S. No	Feature Name
1	Duration	22	Is_guest_login
2	Protocol type	23	Count
3	Service	24	Serror_rate
4	Src_byte	25	Rerror_rate
5	Dst_byte	26	Same_srv_rate
6	Flag	27	Diff_srv_rate
7	Land	28	Srv_count
8	Wrong_fragment	29	Srv_serror_rate
9	Urgent	30	Srv_rerror_rate
10	Hot	31	Srv_diff_host_rate
11	Num_failed_logins	32	Dst_host_count
12	Logged_in	33	Dst_host_srv_count
13	Num_compromised	34	Dst_host_same_srv_count
14	Root_shell	35	Dst_host_diff_srv_count
15	Su_attempted	36	Dst_host_same_src_port_rate
16	Num_root	37	Dst_host_srv_diff_host_rate
17	Num_file_creations	38	Dst_host_serror_rate
18	Num_shells	39	Dst_host_srv_serror_rate
19	Num_access_shells	40	Dst_host_rerror_rate
20	Num_outbound_cmds	41	Dst_host_srv_rerror_rate
21	Is_hot_login		

Information calculation (D): Information theory gave rise to the information gain criteria. The basic principle of information theory is that the information communicated by a message is proportional to the probability and may be quantified in bits as the probabilities minus the logarithm of base 2. Let's say we have a dataset D that has q classifications C1,...Cn. Assume we have a potential test x with m results, which divides D into m subsets D1,...,Dm. Because we only do binary split, m=2 for a number attribute. The chance that one record is chosen from a collection D of data records and announced as belonging to a class Cj is given by,

$$\sum_{j=1}^{m} \left[\frac{freq(C_j,D)}{|\,D\,|} \right] \qquad (4)$$

Where freq (Cj, D) is the number of data records (points) in D for the class Cj, and |D| is the total number of data records in D. As a result, the information conveyed is

$$-\ log_2\left[\frac{freq(C_j,D)}{|D|} \right] \text{ bits} \qquad (5)$$

Summation is conducted across the classes in proportion to their frequencies in D to obtain the anticipated information needed to identify the class of a data record in D before partitioning occurs, yielding Information Calculation (F)

$$\text{Info(D)}=-\sum_{j=1}^{m} \left[\frac{freq(C_j,D)}{|D|}\right] log_2\left[\frac{freq(C_j,D)}{|D|}\right] (6)$$

Assume that the dataset D has been partitioned according to the m results of the test x. after partitioning, the anticipated quantity of information needed to identify the class of a data record in D may be determined as the weighted sum over the subsets, as:

$$\text{Info (F)} = \sum_{i=1}^{n} \left[\frac{|F_i|}{|F|}\right] * info\ (F_i)(7)$$

where | Fi| is the number of data records in the partitioned subset Di after partitioning. IGR (Income Gap Ratio) Calculation (Ai): The following is the information acquired as a result of the partition:

$$\text{Gain (Ai)= info(D)- info(F)} \qquad (8)$$

Clearly, maximization of gain is required. To divide the current data, the gain criteria are used to choose the test or cut that maximizes the gain.

$$\text{IGR } (A_i) = \left[\frac{Info(D)-Info\ (F)}{Info(D)+\ Info\ (F)}\right] * \ 100 \quad (9)$$

The following table summarizes and tabulates the performance and time analysis for various types of assaults. Tables illustrate the detection accuracy and computation time obtained by applying current and proposed feature selection approaches to the characteristics of the KDD'99 Cup data set. The classification of distinct sorts of assaults begins with rule-based classification. Table 2 shows the results of the performance analysis in terms of accuracy and time spent to categories the assaults using the Rule Based Classifier. The detection accuracy and time taken for 5000 records are shown in this table.

Table 2. Performance analysis for Rule based Classification

Attacks	Detection Accuracy (%)	Time in seconds
Dos	76.2	228
Probe	87.3	218
U2R	47.17	227
R2L	64.68	221

Figure 1 depicts the classification time required for rule-based categorization.

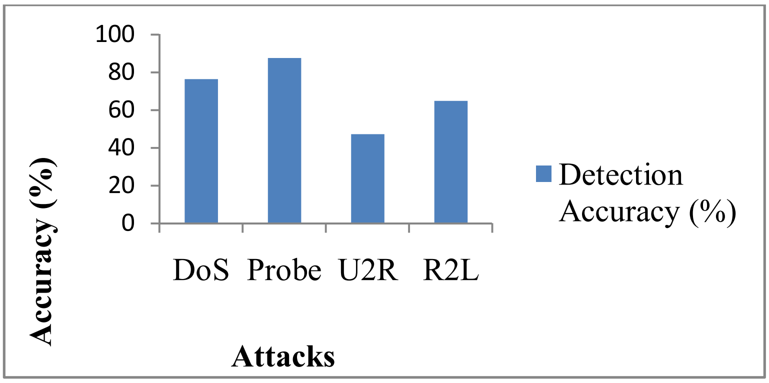


Fig1. Accuracy analysis for Rule based Classification

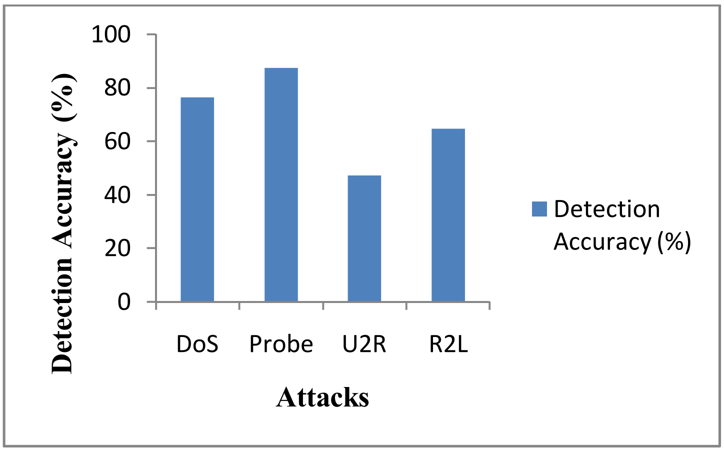
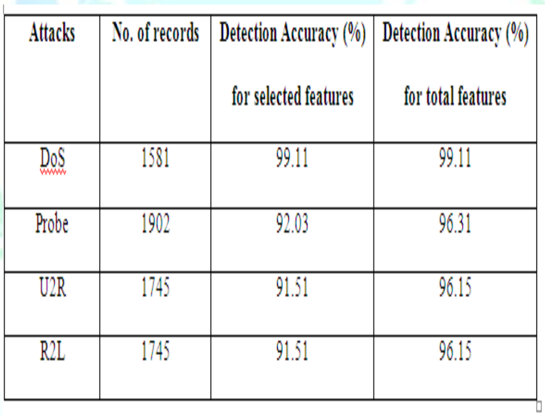


Fig 2. Accuracy analysis for SVM Classification

Figure 2 depicts the classification accuracy for rule-based classification.SVM is used to further classify the records that were collected through Rule Based Classification. Table 3 shows the comparison of detection accuracy for categorization using all 41 features from the KDD cup data set and those chosen using OFS.

Table3. Classification Accuracy using SVM



Attacks	No. of records	Detection Accuracy (%) for selected features	Detection Accuracy (%) for total features
DoS	1581	99.11	99.11
Probe	1902	92.03	96.31
U2R	1745	91.51	96.15
R2L	1745	91.51	96.15

Table 4 shows the time analysis for DoS attacks using SVM classification for variouscategories of 1581 records that were collected..

Table 4. Time Analysis for DoS attack in SVM

Exp No.	Accuracy (%)	
	Selected features using OFS (10)	Total features (41)
1	2.22	2.31
2	2.14	2.26
3	2.03	2.23

4	2.01	2.18
5	2.0	2.17
Avg	2.08	2.23

When compared to the classification done using the 41 features of the KDD cup data set, it is observed that classifying the DoS attack in SVM using features selected in OFS takes less time. The computation time for SVM classification algorithm is identifying DoS attacks and probe attackare depicted graphically in figure 3 and figure 4.

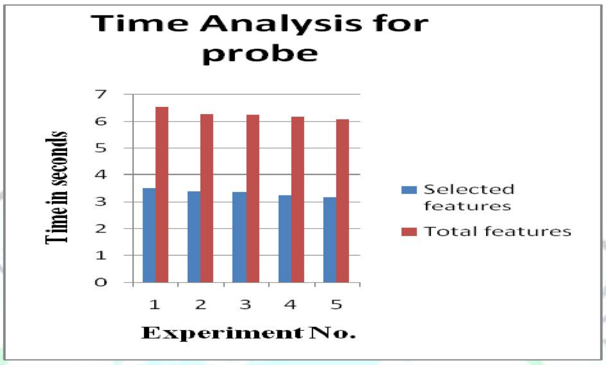


Figure 3. Computation time of DoS attack

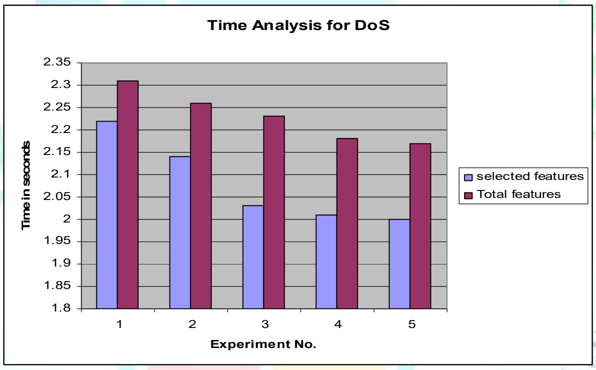
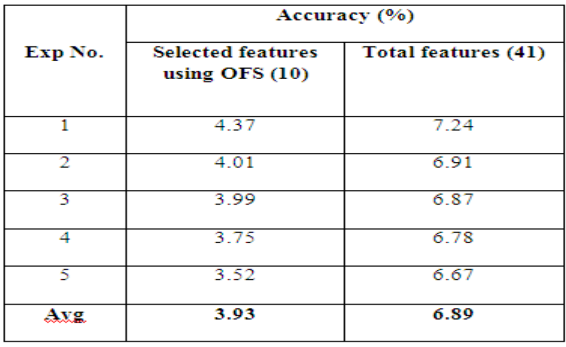


Figure 4. Computation Time of Probe Attack

Table 5 shows the computation time analysis to classify the records obtained from SVM classification as probe attacks.

Table 5. Time Analysis for probe attack in SVM



Exp No.	Accuracy (%)	
	Selected features using OFS (10)	Total features (41)
1	4.37	7.24
2	4.01	6.91
3	3.99	6.87
4	3.75	6.78
5	3.52	6.67
Avg	**3.93**	**6.89**

When compared to the classification done using the 41 features of the KDD cup data set, the time taken to classify the probe attack in SVM using the features selected in OFS takes less time.The calculation time required in WEKA to categorise the probing attack is depicted graphically in fig4.Table 6 shows the time it took to classify the records obtained from rule-based classification as U2Rattack.

Table 6. Time Analysis for U2R attack using SVM

Exp No.	Accuracy (%)	
	Selected features using OFS (10)	Total features (41)
1	3.51	6.52
2	3.41	6.25
3	3.37	6.23
4	3.26	6.17
5	3.17	6.07
Avg	3.34	6.24

When compared to the classification done using the 41 features of the KDD cup data set, it is observed that classifying the U2R attack in WEKA using the features selected in OFS takes less time. The time analysis for 1745 records is shown in table 5. The computing time required in WEKA to categories the U2R assault is depicted graphically in fig5.

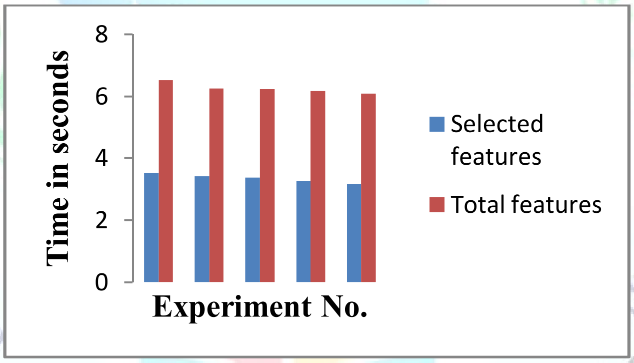
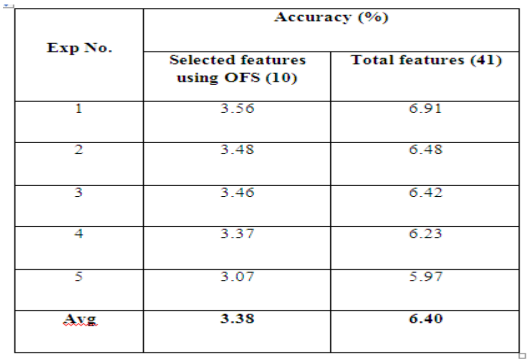


Figure 5. Computation time for classification of U2R attack

Table 7 shows the time it took to categories the 1745 data that were received from rule-based classification as R2Lattack.

Table 7 Time analysis for R2L attack using SVM



Exp No.	Accuracy (%)	
	Selected features using OFS (10)	Total features (41)
1	3.56	6.91
2	3.48	6.48
3	3.46	6.42
4	3.37	6.23
5	3.07	5.97
Avg	**3.38**	**6.40**

When compared to the classification done using the 41 characteristics of the KDD cup data set, it is found that categorizing the R2L attack in WEKA using the features selected in OFS takes less time.The calculation time for categorizing the U2R assault in SVM is depicted graphically in figure 6.

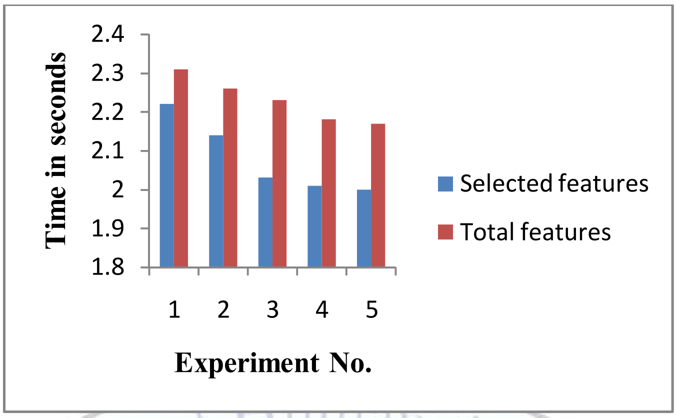
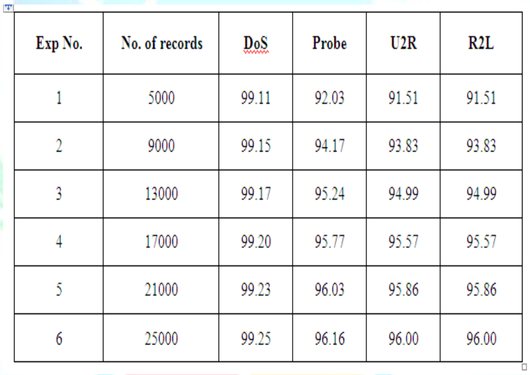


Figure 6. Computation time for classification of R2L attack

Table 8 shows the computation time for SVM for various types of data. As the number of records grows, the accuracy of detecting records with selected characteristics using OFS rises progressively until it approaches the accuracy of recognizing records with all features using SVM.

Table 8. Accuracy Analysis using SVM



Exp No.	No. of records	DoS	Probe	U2R	R2L
1	5000	99.11	92.03	91.51	91.51
2	9000	99.15	94.17	93.83	93.83
3	13000	99.17	95.24	94.99	94.99
4	17000	99.20	95.77	95.57	95.57
5	21000	99.23	96.03	95.86	95.86
6	25000	99.25	96.16	96.00	96.00

The suggested system's installation and outcomes are examined in this paper. The following section will give a conclusion to this study as well as recommendations for future effort.

4. Conclusion:

In this paper, a new intrusion detection model is proposed and implemented for system security by integrating an Optimal Feature Selection method with two classification algorithms. The computation time for identifying and classifying records utilizing all forty-one features of the KDD'99 cup data set is found to be rather lengthy. Only the most essential features are selected by the proposed feature selection method, which helps to reduce the time it takes to identify and classify data. Additionally, the rule-based classifier and SVM aid in improving accuracy. The suggested IDS's major benefit is that it decreases false positive rates while simultaneously reducing calculation time.

REFERENCES:

1. [1] Sindhu, .S, Geetha, S. and Kannan, A. “Decision Tree based Light Weight Intrusion Detection using a Wrapper Approach”, Expert Systems with Applications, Vol. 39, pp. 129–141, 2012.
2. Farid D.M, Jerome Dormont, Nouria Harbi, Nguyen HuuHoa and Rahman, M.Z. “Adaptive Network Intrusion Detection Learning: Attribute Selection and Classification”, International Conference on Computer Systems Engineering, Version 1, pp. 321-337, 2010.
3. Wei Wang, Xiangliang Zhang, Sylvain Gombault and Svein J. Knapskog, “Attribute Normalization in Network Intrusion Detection”, 10th International Symposium on Pervasive Systems, Algorithms, and Networks, pp. 543-559, 2009.
4. Senthilnayaki Balakrishnan, Venkatalakshmi, K and Kannan, A ‘Intrusion Detection System Using Feature Selection and Classification Technique’, International Journal of Computer Science and Application, vol.3, no.4, pp.146-151, 2014.
5. Sarasamma S., Zhu, Q. and Huff, J. “Hierarchical Kohonen Net for Anomaly Detection in Network Security”, IEEE Transactions on System, Man, Cybernetics, Part B, Cybernetics, Vol. 35, No. 2, pp. 302-312, 2005.
6. Wang Jianping, Chen Min and Wu Xianwen, “A Novel Network Attack Audit System based on Multi-Agent Technology”, Physics Procedia, Elsevier,Vol. 25,pp. 2152 – 2157, 2012.
7. Snehal A. Mulay, Devale, P.R. and Garje, G.V. “Intrusion Detection System using Support Vector Machine and Decision Tree”, International Journal of Computer Applications, Vol.3, pp.975-987, 2010.
8. Daramola O. Abosede, A detunmbi A. Olusola, Adeola S. Oladele,. “Analysis of KDD’99 Intrusion Detection Dataset for Selection of Relevance Features”, Proceedings of the World Congress on Engineering and Computer Science, Vol. I, October 20-22, 2010.
9. Wei Lu, Mahbod Tavallaee, Ebrahim Bagheri, Alia A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set”, Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications, Vol. 97,pp.4244-37641, 2009.
10. Leng J, Valli C, and Armstrong L. “A Wrapper-based Feature Selection for Analysis Large Data Set”, Proceedings of 2010 3rd International Conference on and Electrical Engineering (ICCEE), pp. 167-170, 2010.
11. Senthilnayaki B, Venkatalakshmi K and Kannan A 2015, “Intrusion Detection Using Optimal Genetic Feature Selection and SVM based Classifier”, 3rd International Conference on Signal Processing, Communication and Networking (ICSCN). Pp 1-4, 2015.
12. Devale. P.R, Garje. G.V., Snehal A. Mulay, 2012. “Intrusion Detection System using Support Vector Machine and Decision Tree “, International Journal of Computer (0975 – 8887),

Vol. 3, June 2010.

13. Debar, H., Becker, M. and Siboni, D. "A Neural Network Component for an Intrusion Detection System" , IEEE Symposium on Research in Computer Security and Privacy, pp. 240-250, 1992.
14. Du Hongle, Teng Shaohua and Zhu Qingfang, "Intrusion detection Based on Fuzzy support vector machines", International Conference on Networks Security, Wireless Communications and Trusted Computing, pp. 639-642, 2009.
15. Senthilnayaki, B, Venkatalakshmi, K & Kannan, A 2013, 'An Intelligent Intrusion Detection using Genetic based Feature Selection and Modified J48 Decision Tree Classifier', Fifth International Conference on Advanced Computing, pp. 1-7.
16. Senthilnayaki, B, Venkatalakshmi, K and Kannan, A, 'Intrusion Detection System Using Naïve Bayesian and SVM Classifier', International Journal of Applied Engineering Research, vol.10, no.72, pp.408-413, 2015.
17. Weka software, Machine Learning. "Weka 3–Data Mining with Open Source Machine Learning Software in Java" Machine Learning Group at University ofWaikatoWebsite, http://www.cs.waikato.ac.nz/ml/weka/, accessed May 2012.
18. Stuart Russell and Peter Norvig, "Artificial Intelligence", Pearson Education, 2003.
19. Tan, P.N., Steinback, M. and Kumar, V. "Introduction to Data Mining", Addison Wesley, 2006.
20. G. Mahalakshmi and E.Uma. "Machine Learning Based Feature Selection for Intrusion Detection System in VANET", International Conference on Artificial Intelligence, Network Security and Data Science (IeCAN) 2020.
21. G. Mahalakshmi, E. Uma, M. Aroosiya and M. Vinitha, "Intrusion Detection System using Convolutional Neural Network on UNSW-NB15 Dataset", International Conference on Advances in Parallel Computing Technologies and Applications (ICAPTA 2021), IOS Press 2021.