An Improved Medoid Clustering Algorithm For Intrusion Detection Using Web Usage Mining Technique

Preeti Rathi[1], Dr. Nipur Singh[2]

[1]Research Scholar, Dept. of Computer Science, Kanya Gurukul Campus, Dehradun, Uttarakhand
[2]Professor, Dept. of Computer Science, Kanya Gurukul Campus, Dehradun, Uttarakhand

Abstract-

Intrusion detection is one of the applications of web usage mining. In this application, we find the intrusive or worthless data through mining techniques, determine the user behaviour, i.e. new user or existing user, label data according to the users' requirements and detect networks' known and unknown attacks. There are various models of detection of intrusion. Misuse and anomaly detection are types of intrusion. In anomaly detection, the intrusion is unknown and known in misuse. There are various techniques that we discuss in this paper.

We proposed a novel algorithm for intrusion detection using mining techniques based on the medoids and means clustering algorithm. We also compared the proposed algorithm with existing algorithms with high detection and low false alarm rates to detect known and unknown attacks.

Keywords- Detection Rate (DR), False Alarm Rate (FAR), TP, FP, TN, FN.

1. Introduction

Intrusion means unauthorised access. Intrusion detection identifies computer attacks by observing various records processed on the network.

An intrusion detection system (IDS) may be active or passive. Active IDS helps to block suspected attacks automatically based on predefined existing conditions. This type of IDS is known as a detection and prevention intrusion detection system. On the other hand, passive IDS only observes the suspected activities and reports to the administrator for further action. Three essential features, i.e. easy to use, robust, flexible and scalable, are included in each intrusion detection system.

There are five classifications of the intrusion detection system.

1. Network Intrusion Detection System (NIDS) –NIDS is an intrusion detection system used to determine the traffic from all the devices on the network. Traffic is passed on the sub-network of a network or from known attacks.

2. Host Intrusion Detection System (HIDS) –HIDS runs on independent devices called the host. It tracks the incoming and outgoing packets from devices. If malicious activity occurs in the network, an alert message is sent to the admin.

3. Protocol-based Intrusion Detection System (PIDS) deals with protocols that comprise both client and server machines. It is trying to secure the use of hypertext protocol to monitor the client's request.

4. Application Protocol based Intrusion Detection System (APIDS) – APIDS is a group of servers. It identifies the intrusion by monitoring or interpreting communication over application-specific protocols.

5. Hybrid Intrusion Detection System (HIDS) – HIDS is the group of two or more approaches to the intrusion detection system. It is more effective in comparison to other intrusion detection systems.

There are two-way to develop an algorithm for intrusion detection-

1. Design an algorithm based on features which distinguish between normal and intrusive activities or events.

2. Design an algorithm based on models, together all intrusive and normal events.

There are five categories of attacks classes over a network-**[18]**

a. Dos Attack

b. Probing Attack

c. U2R Attack

d. R2L Attack

e. Normal Attacks

DoS Attack: - A denial-of-service attack is a security event that occurs when an attacker takes action that prevents legitimate users from accessing a targeted computer system, devices or other network resources. For example, smurf, neptune, land, pod etc.

Probing attack:- Probing is an attack in which the hacker or attacker scans or probes a machine or a networking device to determine liabilities or vulnerabilities that may later be exploited to compromise the system using probe software—for example, nmap, portsweep, saint, Satan etc.

U2R attack: -User to Root attack class, the hacker has local accessing permission to a machine, and the hacker tries to get the super privileges on that machine. For example, perl, buffer_overflow, xterm etc.

R2L Attack: -Remote to Local attack user has no account or accessing permission on user machine but tries to access the machine. For example, xlock, phy, spy, ftp_write etc.

Normal Attack:-Normal attack connections are generated by simulating user behaviour. For example, normal.

2. Related Work

There are lots of works done by many authors &researchers in the field of intrusion detection. Detection is performed in labelled as well as unlabeled data over the network. Computer network security involves elements like confidentially, integrity and availability.

Discussed work by authors in this field given below-

R. Venketesan et al. **[1]** discuss intrusion detection methods through data mining techniques to detect known and unknown attacks from datasets. In this paper, the authors also discuss the types of intrusion detection systems and categories of intrusion.

Mahza Mabzool et al. **[2]** discuss intrusion detection system based clustering algorithm, i.e. k-mean algorithm. This intrusion detection system determines the input of clustering algorithms and separates the anomaly behaviour of data. Anomaly detection detects unknown attacks and increases accuracy, and decreases false alarms. In this paper, the authors give a solution through different steps of the pre-processing techniques. The first step of data pre-processing is to separate each field of log files using a comma, and then the next step is to clean unwanted data from datasets. After cleaning, it symbolised some long strings to perform calculations fast and recognise the unique user through IP.

Author Kamini Maheshwar **[3]** discusses data mining based intrusion detection techniques. There are two types of intrusion detection, i.e. misuse and anomaly detection. Association rule mining, clustering and classification techniques of data mining used to determine intrusion detection. In this paper, the author discusses some security aspects and clustering techniques to detect intrusion.

Author Subaira A.S. **[4]** surveys the network intrusion detection system based on data mining techniques. In this paper, the genetic algorithm, SVM algorithm, and K-nearest algorithms are discussed by the author and compare these algorithms used to implement the intrusion detection system. IDS is a security system to secure data from different types of attacks. The author also discusses the merits and demerits of existing algorithms. It resolves the optimisation problem and achieves high accuracy and relationship between the dependent and independent variables and speed but needs high storage capacity. The figure below shows the architecture of data mining based on IDS.

Uttam Kumar Dey **[5]** discusses an unsupervised learning approach for network intrusion detection with unlabelled data. In this paper, the author uses unlabelled data set to discuss existing clustering algorithms, k-mean and expectation-maximisation with a high detection rate and low false alarm rate. For the performance evaluation confusion matrix used, two classes are considered actual and predicted classes in this matrix. The confusion matrix contains anomalies and normal TP, FP, TN, and FN data. The NSL-KDD network data set is used to implement a network intrusion detection system.

Author Nadya EI MOUSSAID **[6]** proposed an improved k- mean algorithm to overcome the limitation of an existing k-mean algorithm. For experimental results and comparison of an improved and existing algorithm, KDD cup 99 data set is used. There are four types of attacks: Dos attack, R2L attack, Probe attack, and U2R attack. In this paper improved k-mean algorithm is detecting more than 90% for Dos and R2L attacks and more than 60% for Probe and U2R attacks.

Author Priyanka V. Patil **[7]** discusses pre-processing the weblogs for web intrusion detection. This paper proposed a web intrusion detection system using misuse and anomaly detection using web server logs. Firstly, collect the data from server logs and apply pre-processing techniques to remove unwanted data from the logs dataset. In the detection of intrusion phase, detect the user behaviour, build a model as normal behaviour, and compare this behaviour with abnormal behaviour as an intrusion.

Author Mahesh Malviya **[8]** discusses a review paper titled improving security by predicting anomaly users through web mining. The author proposed a framework for intrusion detection. This paper collects the data from log files and applies the pre-processing technique to remove unwanted data and find the normal or intrusive data. Log files are unstructured format then structured, cleaning data and further classified as a suspect, attacker or normal file.

Author Bhagyashree Deokar **[9]** analyses the drawbacks and advantages of the existing intrusion detection system. The author also proposed an intrusion detection system to minimise the drawbacks of the existing intrusion detection system. This system uses log files and reinforcement learning to detect unknown attacks by reducing the false alarm rate. Reinforcement learning helps to find out the unknown attacks. In this paper author also discuss the flow of finding unknown attacks from log files.

Author M.Deepa **[10]** surveys a comparative study of perceiving intrusion using data mining techniques. This paper discusses the various existing intrusion detection system techniques based on data mining techniques, and the existing system classifies the following parameters: accuracy, detection rate, false alarm, etc. This paper consists of surveys of various existing systems, the methodology used, and the existing system's limitations.

Author Bukola A. Onyekwelu **[11]** discusses pre-processing techniques on university web server log files for intrusion detection. This paper collects log files from different web servers, applies data cleaning to remove unwanted data, and then uses a session identification algorithm to identify each user, i.e. existing or new user. Data discretisation is a process of quantising continuous attributes.

Author Mohsen Eslamnezhand **[12]** proposed intrusion detection based on the min-max k-mean algorithm. In this paper, the author overcomes the shortage of sensitivity to the initial centre in the k-means algorithm and then increases the clustering algorithm's quality. The author also compared the existing k-mean algorithm and the min-max k-mean algorithm. To compare the existing and proposed algorithms NSL- KDD data sets are used. The proposed algorithm is more efficient than an existing k-mean algorithm with a high detection rate and low false-positive detection rate.

Author Gunupudi Rajesh Kumar **[13]** proposed an improved k-mean algorithm for intrusion detection using Gaussian function. This paper designs and analyses the suitability of Gaussian function similarity for intrusion detection. For distance measure, we collect the data from two sources i.e. DARPA and KDD data set and apply the k-mean algorithm to distance metrics; proposed improved k-mean algorithm similarity, fixed the lower and upper bound for metrics.

Author K.S. Anil Kumar **[14]** discusses the various clustering algorithm for intrusion detection. In this paper, the author compares the performance of the clustering algorithm using the DARPA data set. K- Medoid and improved k-mean clustering with optimum cluster centroid initialisation algorithm. This algorithm achieves high accuracy. Clustering algorithms are used in an intrusion detection system to separate normal and abnormal behaviour.

Author S. Revathi **[15]** focuses on a detailed analysis of NSL-KDD dataset using various machine learning techniques for intrusion detection. This data set includes 41 features. This paper discusses various classification algorithms and compares them using KDD data sets. Performance measure parameters accuracy and false alarm rate are used with high accuracy and low false alarm rate. For comparison, KDD and DARPA data set was used.

Author Kapil Wankhede **[19]** proposed an efficient approach for intrusion detection using data mining techniques. The main objective of this paper is to improve the detection rate and decrease the false alarm rate. In this paper, the author discusses the hybrid data mining approach, including features such as section, filtering, and clustering. Intrusion detection systems are those systems which detect intrusion, i.e. unauthorised access, with achieved high accuracy and low false alarm rate.

3. Data Set Description

We collect the data from the UCI repository. KDD Cup **[16]** data will be considered to detect intrusion. NSL_KDD dataset is also used to detect intrusion over the network. KDD Cup data set contains 41 attributes over approximately 4, 94,003 data instances. The dataset comprises labelled and unlabelled records. There are various types of attacks simulated into four categories, i.e. DOS attack (Denial of Service), Probe attack, R2L attack (Remote to User) and U2R attack (User to Remote). Some attacks are considered normal attacks generated by user behaviour. The total number of attacks is 34, which belongs to a different category.

There are four protocols which are considered in KDD Cup data set, i.e. Transmission Control Protocol (TCP), User Datagram Protocol (UDP), Internet Control Management Protocol (ICMP), Internet Group Management Protocol (IGMP) etc. TCP protocol is affected by more than five attacks compared to other protocols because it handles all transmission requests and is a connection-oriented protocol.

UDP protocol affected fewer attacks than another protocol in KDD Cup data set because it is a connectionless protocol. ICMP and IGMP protocols are internet protocols. In KDD Cup 10-15% dataset has these protocols. For the implementation of proposed and existing algorithms, we used WEKA tool.

WEKA is a tool for data mining and machine learning implemented at the University of Waikato in New Zealand in 1977 **[17].** WEKA software is programmed in JAVA language, and it has a GUI interface to interact with data files. WEKA supports data sets in the form of ARFF. WEKA can expand and contain a new algorithm for machine learning in it. There are two categories of the dataset, i.e. training and testing, given below. We used 40% of the whole dataset, i.e. and 25% data set used for training and 15% for testing dataset.

Training Dataset- A training set is a dataset used to train the model and pick the specific features from the training dataset.

Test Dataset- The test set is a dataset used to measure how well the model performs at predicting on that dataset.

The below table shows the classification of attacks with instances of dataset & classes:

Table- 1 Classification of Attacks with instance & class

S.No	Class of Attack	Instances	Types of Attack	% age
1.	DoS	391468	smurf, neptune, land, pod	79.2

2.	Probe	4104	ipsweep, mscan, nmap, portsweep	0.82
3.	U2R	52	Buffer_overflow, perl, xterm,	0.01
4.	R2L	1115	ftp_write, imap, multihoop	0.21
5.	Normal	97264	Normal	19.6

The table below shows the classification of classes with training & testing data set for implementing existing and proposed algorithms. We collect 40% of the whole data because the data size is huge, and 25% for training, 15% for testing-

Table- 4.3 Classification of classes with training & testing data

	Normal	DoS	Probe	R2L	U2R	Total	%
Training	14619	32246	2406	89	21	49401	25%
Testing	5017	22450	2104	48	21	29640	15%

4. The flow of Proposed Work

In this part of the paper, we discuss the implementation of an intrusion detection algorithm for unlabeled data. Firstly, we collect the data set from the UCI repository. We used 40% of the KDD Cup data set because this dataset contains many labelled and unlabeled data. We consider the unlabeled dataset and then apply the Apriori and FP tree algorithm for labelling the data. There are two methods or processes to extract a valuable attribute from whole attributes, i.e. filtering and wrapping. We used the filtering process because it treats missing value as a separate value using the **Info_Gain** attribute in WEKA. It evaluates the worth of an attribute by measuring the information gained concerning the class, and we get 21 attributes from 41 attributes. After filtering, retrieves useful values, these values are used to find intrusive or normal data.

Info_Gain (class, attribute) =H (class) -H (class|attribute)

Then discretisation process is used to convert real-valued into ordinal or categorical attributes; after that dataset is divided into training & testing data for further processing. We train the dataset for applying a classifier and clustering existing algorithm to evaluate performance metrics for comparison with a proposed algorithm with a trained dataset. Implement the proposed algorithm named the IMCID.

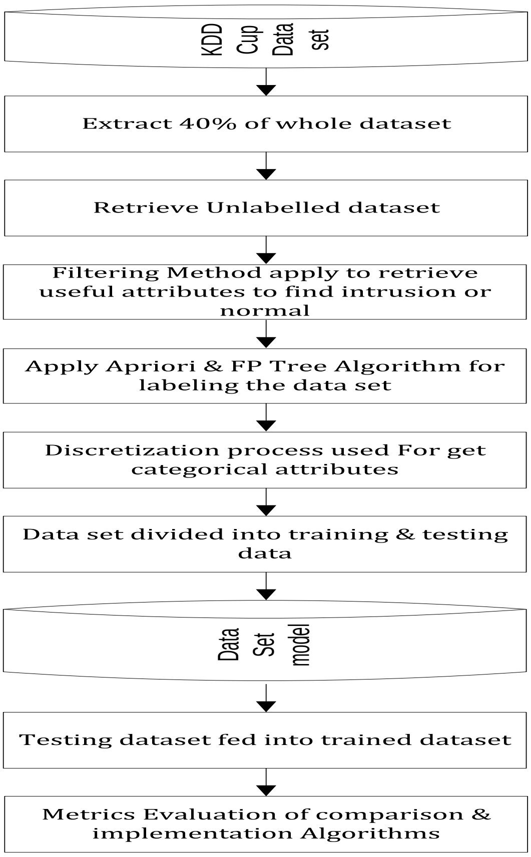

KDD Cup Data set
Extract 40% of whole dataset
Retrieve Unlabelled dataset
Filtering Method apply to retrieve useful attributes to find intrusion or normal
Apply Apriori & FP Tree Algorithm for labeling the data set
Discretization process used For get categorical attributes
Data set divided into training & testing data
Data Set model
Testing dataset fed into trained dataset
Metrics Evaluation of comparison & implementation Algorithms

Figure- 1 Flow of Detection of Intrusion

5. Metrics to measure the performance of the system

Improving the performance of the intrusion detection system requires a high detection rate and low false alarm rate. The following terms are used to calculate the performance of the intrusion detection system.

5.1 Accuracy is computed as the ratio between correctly detected attacks and the total number of attacks. It is also known as the true positive rate (TPR), the probability of true detection.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FN+FP}}$$

5.2 TP - Intrusions that are successfully detected by the ID, i.e. the attack data is classified as an attack.

5.3 FP- Normal/non-intrusive behaviour, i.e., wrongly classified as intrusive by the IDS.

5.4 TN- Normal/non-intrusive behaviour that is successfully labelled as normal/non-intrusive by the IDS.

5.5 FN- Intrusions that are missed by the IDS and classified as normal/non-intrusive.

5.6 Recall- It is also known as True Positive Rate (TPR), sensitivity or detection rate. It is used to calculate the performance evaluation of IDS.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{FN+TP}}$$

5.7 Precision is calculated by the false positive and true positive instance, i.e. predict positivity over the total dataset.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$$

5.8 FAR- It is also known as the false-positive rate. It is a probability of false detection; and is computed as the ratio between the number of wrongly detected attacks and the total number of attacks.

$$\boldsymbol{FalsePositiveRate} = \frac{\boldsymbol{FP}}{\boldsymbol{TN+FP}}$$

5.9 F-score- It is considered as a harmonic mean between recall and precision. F-score provides balancing between recall and precision metrics.

$$F-score=\frac{2*P*R}{(P+R)}$$

Where P=Precision, R= Recall.

5.10 Confusion Matrix- A confusion matrix is a matrix that represents the classification result. It represents true and false classification results.

Table 2:- Confusion matrix

Actual	Predicted	Predicted
	Attack	Normal
Attack	TP	FN
Normal	FP	TN

6. Algorithm for detection of intrusion

Proposed a named as IMCID is used to detect intrucive or normal data and achieve high accuracy and low false alarm rate. To calculate the distance between clusters average distance formula is used.

A new association rule mining algorithm, FP growth algorithm, is used to overcome the apriori algorithm and find the frequent pattern from the data set. FP growth algorithm is an improved apriori algorithm. This algorithm helps to generate rules or patterns.

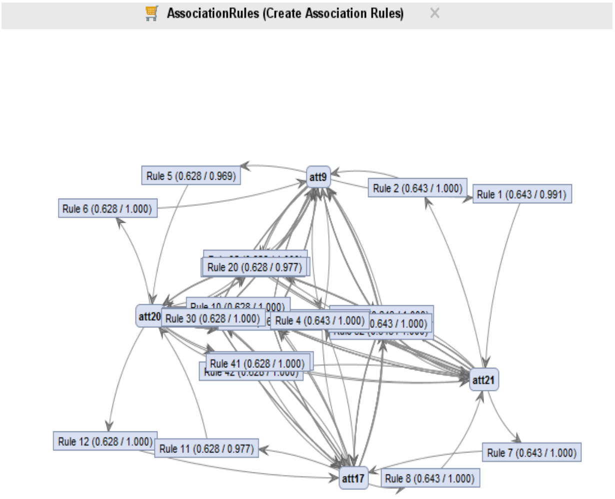


Figure- 2 FP Tree generated from the training dataset

To minimise the drawbacks of existing k-mean and k-medoid and high detection rate & low false-positive rate we proposed an algorithm named Improved Medoid Clustering Algorithm for intrusion detection. K-Medoid algorithm based on clusters' partitioning technique, contains x objects. In k-medoid algorithm, the different initial sets of medoid lead to different result clustering. The initial centroid and initial medoid are selected randomly in K-Mean and k-medoid algorithms. **Data Normalisation-** Data normalisation is the method to normalise the data according to a specific value. There are two methods of data normalisation- Min-Max normalisation and Z-score normalisation.

Min – Max Normalization- In this normalisation technique, subtract each attribute's minimum value and then divide the difference by the range of attributes. The below formula shows min-max normalisation-

$$Nor\_Va=\frac{v-min(A)}{max(A)-min(A)}\big(n\_max(A)-n\_min(A)\big)+n\_min(A)$$

Where min (A) & max (A) are the minimum and maximum values of an attribute A. v is the old value of each entry from the dataset, and after normalisation, Nor__Va is the normalised value. n_max and n_min are boundary value of range required, i.e. [0, 1] initial value.

Z-Score Normalization-It is also called zero score normalisation. This technique is based on mean and standard deviation. The formula of z-score normalisation is-

$$\boldsymbol{Nor\_Va = \frac{v - mean}{S.D.}}$$

Where v is the old value of each entry from the dataset, and after normalisation Nor__Va is the normalised value.

There are two phases of improved medoid clustering algorithm for intrusion-

Phase I

1. Collect the data set from KDD Cup where data is labelled & unlabelled.
2. Using min-max normalisation, apply data normalisation to normalise data with a specific range, i.e. [0, 1].
3. Convert unlabelled data into labelled data using a hybrid approach, i.e., density-based and hierarchical clustering algorithm.
4. Retrieve a high-quality dataset to detect intrusion from the labelled dataset in phase II.

Phase II

1. We used a labelled data set which is collected from phase I.
2. For improved medoid clustering algorithm for intrusion detection, both k-medoid and its improvement are used.
3. We get an improved detection rate with a low false-positive rate.

Algorithm-

Improved Medoid Clustering Algorithm for Intrusion Detection (IMCID)

```
INPUT- Dataset with n objects (trained_dataset) previous_comp_value
OUTPUT- intrusive_Data and
        Normal_Data
[1] Nor_Va (min (A), max (A))
//Normalise the dataset, remove duplicate data and minimise the dataset for detection of intrusion using Min-Max normalisation
[2] T=Random(trained_dataset)
//Select the initial medoid randomly from trained_dataset.
[3] Retrieve normalised dataset and calculate dissimilarity measures using average
        Distance formula-
$distance_{avg}\,(i,j) = (\frac{1}{N}\sum_{x=1}^{N}(p_{ix} - q_{jx})^2)$
Where, i=1, 2, 3,...,K, j=1, 2, 3.....K, and N be the objects in cluster X, and p & q are randomly selected points from data.
//This method also detect known attacks.
If attacks are unknown, then use the cosine function to calculate distance.
$Cosine(i,j) = \frac{\sum_{x=1}^{N} p_{ix}\, q_{jx}}{\sqrt{\sum_{i=1}^{n} {p_{ix}}^2}\sqrt{\sum_{i=1}^{n} {q_{jx}}^2}}$
[4] Collections.sort(X)
// Arrange the value of objects in cluster X in ascending order and again choose the initial medoid with minimum value.
[5] Associate each object to its nearest medoid & calculate the optimal_value
//optimal_value as a sum of the value of each object with its medoids.
[6] Swap the medoid values to get the minimum value using the objective function for each pair of selected and non-selected objects & calculate the total swapping cost for all objects.
$TSC_{mn} < 0$
Where m & n are selected and non-selected object.
[7] Repeat step [6] and continue till m is replaced by n.
[8] Again associate each objects with nearest medoid and compute new_value as in step [5].
[9] If (new_value == previous_comp_value)
        {
            Retrieve Normal_Data
        }
        Else
        {
            Retrieve Intrusive_Data
        }
```

7. Experimental Results &Evaluation

In our research, we collect the dataset from the UCI repository. We collect the whole dataset and then divide the dataset into training and testing datasets. For simulation, we used the WEKA tool. We proposed an algorithm with a high detection rate and low false alarm rate. Table 3 shows the experimental result of the proposed algorithm.

To compare the proposed algorithm with the existing algorithm, we achieved the following parameters: accuracy, false alarm rate, and detection rate. Accuracy is defined in percentage; it means how much better accuracy we achieved from the existing algorithm. FAR decides the probability of false detection.

Figure 3 shows the accuracy percentage for existing algorithms with the proposed algorithm. We used k-mean, improved k-mean, k-medoid, and improved k- medoid for comparison.

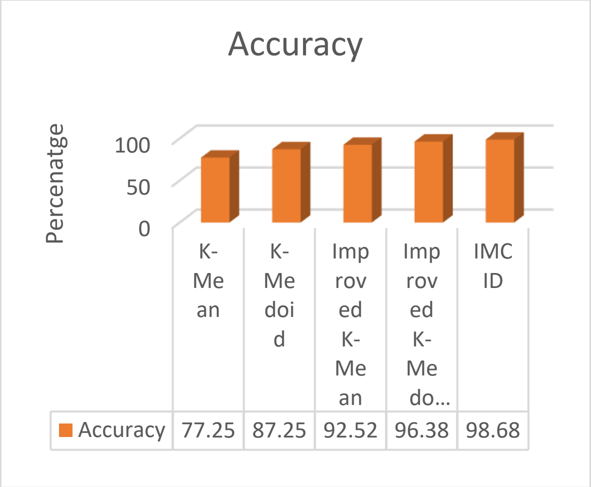


Figure 3 Comparison of existing algorithms with proposed IMCID algorithm with parameter accuracy

Below table 4 shows the experimental result of the proposed algorithm with the existing algorithm with parameters such as detection rate and false-positive rate.

Table- 4 Comparison of existing algorithms with proposed IMCID algorithm with parameter detection rate & FAR

Algorithms → / Parameters ↓	K-Mean	K-Medoid	Improved K-Mean	Improved K-Medoid	IMCID
Detection Rate	82.3	85.25	89.82	91.21	95.68
FAR	5.2	4.9	4.5	3.2	2.8

Below, figure 4 shows the graphical representation of detection rate & false alarm rate with the comparison of existing algorithms & IMCID algorithm. To enhance the detection rate, we used an efficient method to select medoid and give better results and resume the problem detection of intrusion from the dataset.

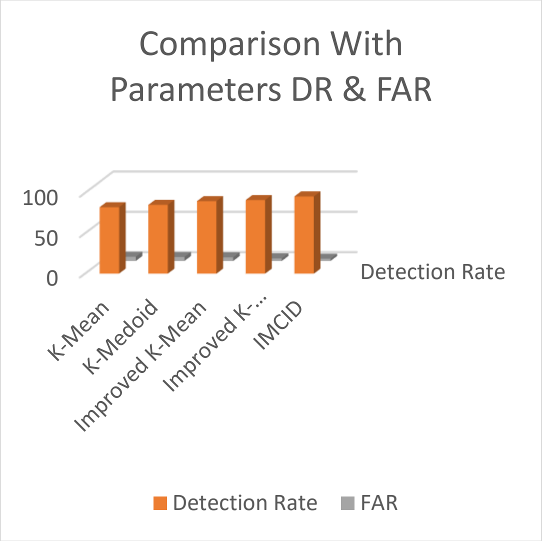
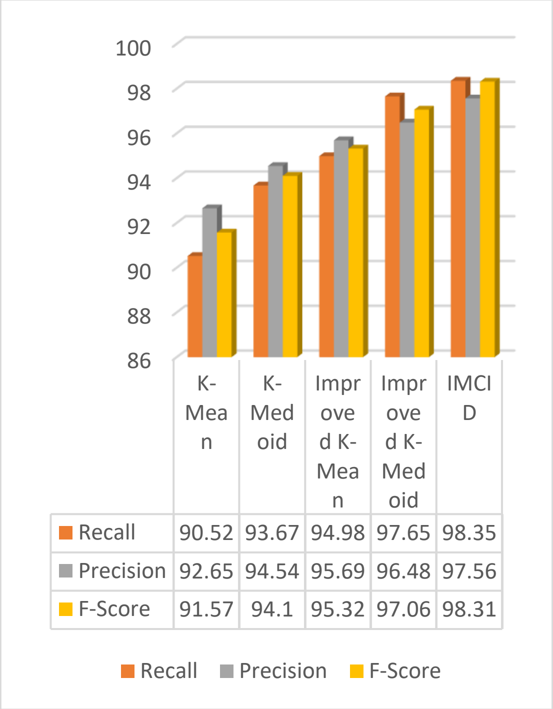


Figure 4 Graphical Representation of existing algorithms with IMCID with parameters DR & FAR

Table - 3 Experimental Result of Metrics

Metrics			**Types of Attacks**							
	Normal		**DOS**		**PROBE**		**R2L**		**U2R**	
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**	**Training**	**Testing**	**Training**	**Testing**
TN	9423	3241	19461	11423	1248	1037	58	33	12	12
TP	5132	1715	12653	10897	1133	952	28	12	7	7
FN	22	19	59	47	16	12	1	1	1	1
FP	42	42	73	83	9	13	1	1	1	1
Recall	99.5	98.9	99.7	99.5	98.6	98.7	96.5	92.3	87.5	87.5
Precision	99.3	97.6	99.4	99.2	99.2	98.6	93.3	92.3	87.5	87.5
Accuracy	99.5	98.7	99.5	99.4	98.9	94.5	92.1	91.6	90.4	90.4
F-Score	98.9	98.2	99.5	99.3	98.8	98.6	94.8	92.3	87.5	87.5

Figure 5 shows the comparison in graphical form of the proposed algorithm IMCID with the existing algorithm with recall, precision & F-score.



	K-Mean	K-Medoid	Improved K-Mean	Improved K-Medoid	IMCID
🟧 Recall	90.52	93.67	94.98	97.65	98.35
⬜ Precision	92.65	94.54	95.69	96.48	97.56
🟨 F-Score	91.57	94.1	95.32	97.06	98.31

Figure 5 Comparison in graphical form of existing algorithms with proposed IMCID algorithm with parameter Recall, Precision & F-Score

Table 5 shows the confusion matrix of the known attacks, and table 6 shows the confusion matrix of the unknown attacks from the dataset using the average distance & cosine method. Find intrusion and normal data from the set if known and unknown attacks are found. The proposed algorithm gives a better result for finding intrusion from the dataset.

Table 5:- Confusion Matrix for Known Attack

Predict ➡ **Actual** ⬇	**Attack**	**Normal**
Attack	90.52	93.67
Normal	92.65	94.54

Table 6:- Confusion Matrix for Unknown Attack

Predict ➡ **Actual** ⬇	**Attack**	**Normal**
Attack	68.64	71.59
Normal	69.45	64.68

Table 7 shows the intrusive and non-intrusive data, and the proposed algorithm gives better results in comparison to the existing algorithm.

Table 7:- Intrusive and Non-Intrusive Dataset

Labelled Dataset ➡ **Algorithms** ⬇	**Intrusive Data (12681)**	**Non-Intrusive Data (25685)**

K-Mean	10895	21558
K-Medoid	10998	21998
Improved K- Mean	11258	23548
Improved K- Medoid	11998	24996
IMCID	12584	25485

Table 8 shows the detection rate of various attacks with existing and proposed algorithms. Our proposed algorithm gives a better detection rate for each attack. The existing algorithm is not efficient in finding remote-to-login and user-to-root attacks.

Table 8:- Detection Rate of various attacks

Attacks	**Detection Rate**				
	K-Mean	**K-Medoid**	**Improved K-Mean**	**Improved K-Medoid**	**IMCID**
Normal	84.65	88.96	91.65	95.69	98.65
DoS	79.87	81.35	86.76	96.76	97.83
Probe	78.54	79.82	84.21	87.87	95.26
R2L	70.87	72.26	74.39	79.65	91.27
U2R	69.74	71.78	73.56	78.96	90.87

Below, figure 6 compares different attacks, i.e., Normal, DoS, Probe, R2L, U2R, and the detection rate in graphical form.

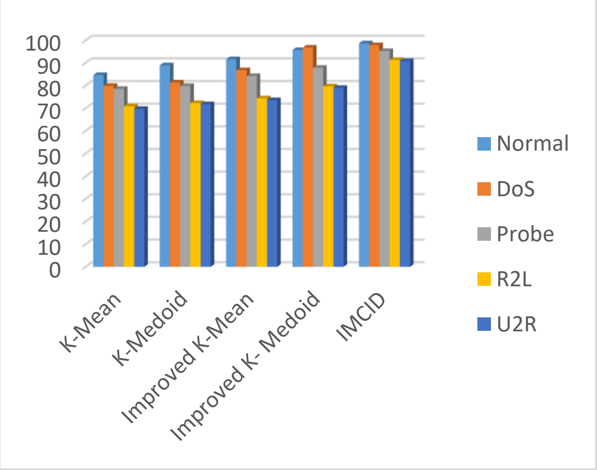


Figure 6 Graphical form of Detection Rate of various attacks

8. Conclusion

Intrusion detection is the process of detecting the abnormal behaviour of the user. In this paper, we proposed an algorithm for intrusion detection based on the medoid algorithm. In existing medoids algorithms, problem to detect the required initial medoid. IMCID algorithm removes this drawback. We also added the method to convert unlabelled datasets to labelled datasets and achieve high-quality data. Using TP, FP, TN, FN, Recall, Precision, Detection Rate, F-Score, and Accuracy & FAR to compare a proposed algorithm with existing algorithms. We also identify known and unknown attacks and find intrusive or normal data. We compare k-mean, k-medoid, improved k-mean, and improved k-medoid algorithms with IMCID.

9. References

[1] R.Venkatesan, Dr. R. Ganesan, Dr. A. Arul Lawrence Selvakumar, "A Survey on Intrusion Detection using Data Mining Techniques", International Journal of Computers and Distributed Systems (IJCDS), ISSN No. 2278-5183, Volume No. 2, Issue No. 1, Page No. 54-58, December 2012.

[2] Mahza Mabzool, Mina Zolfy Lighvan, "Intrusion Detection System Based on Web Usage Mining", International Journal of Computer Science, Engineering and Applications (IJCSEA), ISSN No. 0275-2190, Volume No. 4, Issue No. 1, Page No. 1-8, Feb 2014.

[3] Kamini Maheshwar, Divakar Singh, "A Review of Data Mining based Intrusion Detection Techniques", International Journal of Application or Innovation in Engineering & Management (IJAIEM), ISSN No. 2319-4847, Volume No. 2, Issue No. 2, Page No. 134-142, Feb 2013.

[4] Subaira A.S, Anitha P, "A Survey: Network Intrusion Detection System based on Data Mining Techniques", International Journal of Computer Science and Mobile Computing (IJCSMC), ISSN No. 2320-0881 , Volume No. 2, Issue No. 10, Page No. 145-153, Oct 2013.

[5] Uttam Kumar Dey, Mohammad Alauddin, Tanzillah Wahid, "Network Intrusion Detection with Unlabelled Data using Unsupervised Clustering Approach", International Journal of Engineering Science and Computing (IJESC), ISSN No. 2321-3361 ,Volume No. 9, Issue No. 2 , Page No. 19661-19664 , Feb 2019.

[6] Nadya El MOUSSAID, Ahmed TOUMANARI, Maryam ELAZHARI, "Intrusion Detection Based On Clustering Algorithm", International Journal of Electronics and Computer Science Engineering (IJECSE), ISSN No. 2277-1956, Volume No. 2, Issue No. 3, Page No. 1059-1064, 2018.

[7] Priyanka V. Patil, Dharma raj Patil, "Pre-processing Web Logs for Web Intrusion Detection", International Journal of Applied Information Systems (IJAIS), ISSN No. 2249-0868, Volume No. 4, Issue No. 1, Page No. 11-15, 2012.

[8] Mahesh Malviya, Abhinav Jain, Nitesh Gupta, "Improving Security by Predicting Anomaly User Through Web Mining: A Review", International Journal of Advances in Engineering & Technology (IJAET), ISSN No. 2231-1963, Volume No. 1, Issue No. 2 , Page No. 28-32, May 2011.

[9] Bhagyashree Deokar, Ambarish Hazarnis, "Intrusion Detection System using Log Files and Reinforcement Learning", International Journal of Computer Applications (IJCA), ISSN No. 0975 – 8887, Volume No. 45, Issue No. 19, Page No. 28-35, May 2012.

[10] M. Deepa, Dr. P. Sumitra, "A Comparative Study of Perceiving Intrusion Using Data Mining Techniques", International Journal Of Engineering And Computer Science (IJECS), ISSN No. 2319-7242, Volume No. 5, Issue No. 12, Page No. 19494-19497, Dec 2016.

[11] Bukola A. Onyekwelu, B. K. Alese and A. O. Adetunmbi, "Pre-Processing of University Webserver Log Files for Intrusion Detection", International Journal Computer Network and Information Security (IJCNIS), ISSN No. 2074-9090 , Volume No. 4, Issue No. 1, Page No. 20-30, 2017.

[12] Mohsen Eslamnezhad, Ali Yazdian Varjani, "Intrusion Detection Based on Min Max K-means Clustering", International Symposium on Telecommunications (IST), IEEE, ISBN No. 978-1-4799-5359-2, Page No. 804-808, 2014.

[13] Gunupudi Rajesh Kumar, N Mangathayaru, G Narasimha, "An improved k-Means Clustering algorithm for Intrusion Detection using Gaussian function", International Conference on Engineering and Management Information System (ICEMIS), ACM,

ISBN No. 978-4503-3418-1, Page No. 1-7, Sep 2015.

[14] Dr. K.S. Anil Kumar, Anitha Mary M.O. Chacko, "Clustering Algorithms for Intrusion Detection: A Broad Visualization", International Conference on Information and Communication Technology for Competitive Strategies (ICTCS), ACM, ISBN No. 978-1-4503-3962-9, Page No. 1-4, Mar 2016.

[15] S. Revathi, Dr. A. Malathi, "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection", International Journal of Engineering Research & Technology (IJERT), ISSN No. 2278-0181, Volume No. 2, Issue No. 12, Page No. 1848-1853, Dec 2013.

[16] KDD Cup 1999 Data EB/OL].(1999).http://www.kdd.ics.uci.edu/databases/kddcup.html"

[17] Dash, Ranjita Kumari, "Selection of the Best Classifier from Different Datasets using WEKA", International Journal of Engineering Research and the Technology (IJERT), ISSN No. 2278-0181, Volume No. 2, Issue No. 3, Page No. 2-7 , March 2013.

[18] SACHIN S. PATIL, PROF. DEEPAK KAPGATE, PROF. P.S. PRASAD, "Efficient Concept for Detection of Web Based Attacks Using ID3 Algorithm ", International Journal of Computer Science and Mobile Computing (IJCSMC), ISSN No. 2320-088X, Volume No. 3, Issue No. 5, Page No. 321-323, May 2014.

[19] Kapil Wankhede, Sadia Patka,Ravindra Thool, "An Efficient Approach for Intrusion Detection Using Data Mining Methods", International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, ISBN No. 978-1-4673-6217-7, Page No. 1615-1618, 2013.