# Towards an Effective Feature Selection in NIDS

M. JAMES STEPHEN*, T. AUGUSTINE#, P. VASUDEVA REDDY#, P.V.G.D. PRASAD REDDY+

*Professor, Department of CSE, WISTM Engineering College, Visakhapatnam
#Department of Engineering Mathematics, +Department of CS & SE
A.U. College of Engineering, Andhra University, Visakhapatnam, AP, INDIA
Email address of corresponding author: jamesstephenm@yahoo.com

## ABSTRACT:

Internet is now a common commodity to every common man. The data is consistently on network. Andall the financial transactions are being done online. Hence providing Network and Internet security has becomea serious worldwide issue. Cyber terrorism is getting increased with time. Apart from this, these cyber attacks are causing severe damages to different sectors that include individuals, private and public organizations and governments. This provides the need of developing an effective Network Intrusion Detection Systems (NIDS) to safeguard these attacks. The performance of any NID System depends on the performance of classification algorithm that is employed. The performance of the classification algorithm depends on the effectivenessin feature selection. On these grounds, in order to develop an effective Network Intrusion Detection System or at least to propose some enhancements to the existing Network Intrusion Detection Systems, the basic understanding on the methodology and working process of NIDS is required.

In this paper, some new methodologies towards feature selection and effective evaluation of classification performance are presented. A rigorous study and analysis of the basic concepts in Network Intrusion Detection Systems and fundamental concepts in feature selection are also described. In order to analyze and understand the classification process in NIDS, the oldest and well known machine learning classification algorithm, which is called Support Vector Machine (SVM) is also presented in this paper. The theoretical development of SVM and mathematical basis for the classification algorithms are also explained.Two major feature selection approaches are analyzed. One is the oldest Recursive Feature Elimination (RFE) [4] and other is the latest Recursive Feature addition (RFA) [1]. A new Trigram Technique is developed to handle long payload features. Various experiments were carried out with different feature selection approaches on the benchmark data set ISCX 2012 using SVM classification algorithm. A new composite performance evaluation metric, DAFAR is introduced. Comparisons are made through the empirical results and some problems with the existing system are identified along with important observations and proposed few solutions.

***Keywords:***Network Intrusion Detection System, NIDS, SVM, Feature selection, Intrusion Detection, Recursive Feature Elimination, Recursive Feature Addition, ISCX 2012

## 1. Introduction:

The rise of Internet has dramatically affected the entire scenario of the whole world. The users of internet are exponentially growing. Today, internet has become a part of one’s daily life. From ticket booking to millions of banking transactions, internet is required. The dependency on internet is increasing day by day from every field in all disciplines. Internet brought a paradigm shift in the communication world. However, this growth is not without risks and caused the Internet a target for many types of threats.

There are thousands of cyber crimes being reported every year all over the world. The estimated worldwide of total data lost to cybercrimes range from hundreds of Billions to over one Trillion Dollars and in the last 12 months 77% of businesses reported a data breach. The worst part is that 63 percent of businesses don’t have a grown-up system to track their sensitive data [7]. Studies by IBM Security and Ponemon Institute [8] in 2017 on 419 companies in 13 countries conclude that the average total cost of data contravene is $3.62 million and $141 is the average cost per lost or stolen records. The below figure 1 gives even more clear picture.

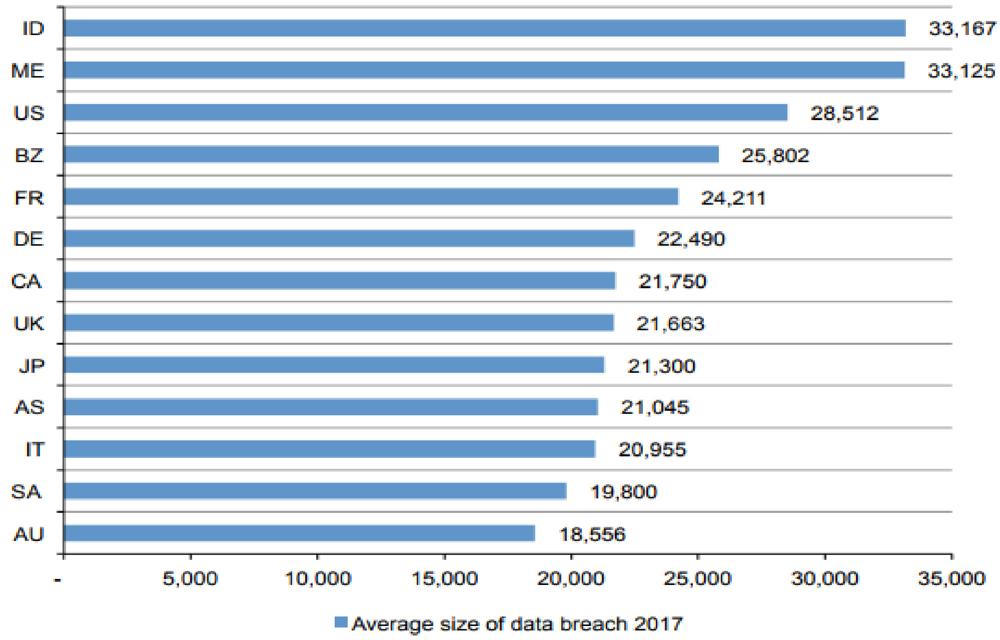


Figure 1: The average number of breached records by country or region

Figure 1 reports the average size of data breaches for organizations in these countries. On average, organizations in India (ID), the Middle East (ME) and the United States (US) had the largest average number of breached records. Australia (AU), South Africa (SA) and Italy (IT) stands on smaller average number of breached records.

It is important to note that 9 of 13 countries demonstrated an increase in the probability of data breach. India was on the largest increase rate at 8.7 %, followed by France at 4.2 %. Figure 2 presents the average cost per data breach over four years and comparing that with FY 2017.

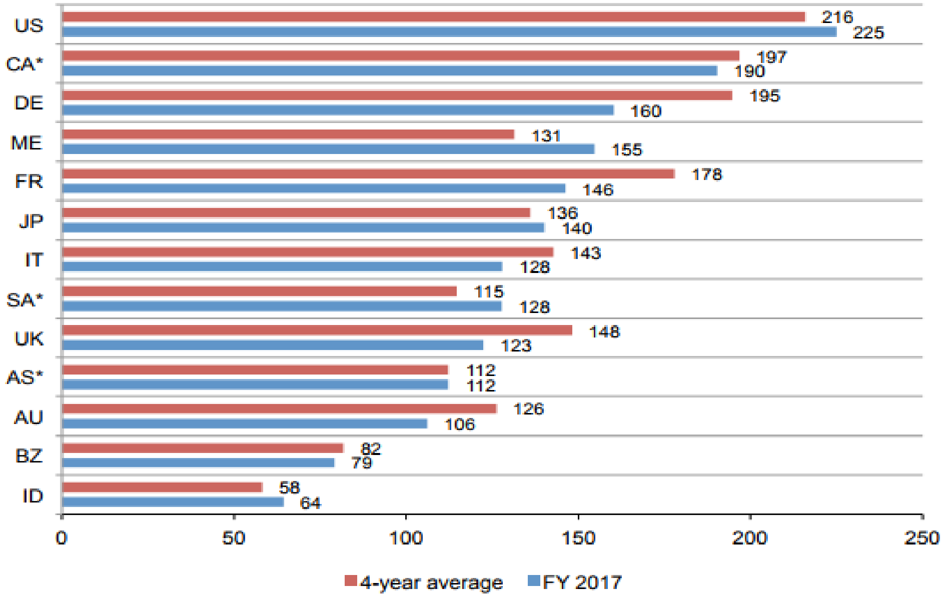


Figure 2: The 2017 cost per data breach compared to the four-year average measured in US $

This gives the motivation to think of a solution for this very challenging problem of providing security to the data on the network. The work in this paper is presented as follows: Section 2 provides some of the basics concepts of NID systems and a detailed theoretical development of Support Vector Machine is explained. The mathematical formulation for classification problem with respect to SVM is elaborated in this paper. Section 3 deals with the challenges in feature selection, Trigram, a new technique to handle long payload features is presented. Two feature selection algorithms RFE and RFA are explored. Section 4 presents the application of Support Vector Machine (SVM) on different datasets from ISCX 2012. A new composite performance evaluation metric, called DAFAR is presented in Section 5. Later in Section 6, various comparative results of the proposed techniques are presented. Finally this paper ends with conclusions and future scope in Section 7.

## 2. Network Intrusion Detection System (NIDS)

As a consequence of the aforesaid damages of intrusions on the Internet, exploring an efficient countermeasure is really a necessity to protect Internet users against these intrusions. The development of these attacks has fascinated many researchers around the sphere to investigate and try a solution for these problems. Therefore, the efforts of network security research community from different parts of the world have resulted in a countermeasure called an Intrusion Detection System (IDS). The IDS is accountable for detecting offensive activities on a system (computer or network) and should raise an alarm in the case of detecting an intrusion. IDSs can be categorized into two types according to their place of operation: Host based IDS and Network based IDS (NIDS).

(a) **A Host-based IDS** (HIDS) resides in a host computer, where it monitors and analyzes log entries for particular information. Its operation requires periodically looking for new log entries and comparing them with pre-defined rules. An alarm should be raised if by HIDS if it finds a match between a log entry and a rule. Some latest versions of HIDS examine all the system calls to look for certain attack signatures. An action will be initiated by the HIDS, in case a match is found between a system call and any of the signatures [9]

(b) **A Network based IDS** (NIDS), on the other hand, involves monitoring all the traffic passing through a network card to and from the network. The NIDS then examines the traffic according to a set of rules and attack signatures, Internet Protocol (IP) and transport layer headers of packets, and even the content of such packets to determine if the traffic contains an intrusion. If it finds one, then an alarm is blown.

In the present work, focus is on NIDS to find out the detection of any intrusion. With this basic information on NIDS, the following section presents the theoretical development of Support Vector Machine.

## 2.1 Support Vector Machine (SVM):

The basicsof Support Vector Machines (SVMs) were established by Vapnik (1198), Burges (1998), [27], based on statistical learning theory, to solve the classification problem. Due to promising empirical performance and many attractive features,it is gaining alot of popularity. In this section, the theoretical development and mathematical formulation is presented

### 2.1.1 Linear SVM:

Linear SVM is based on the concept of linearly separabilityor hyperplane classifier.

Suppose we have N training data points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ where $x_i \in R^d$ and $y_i \in \{\pm 1\}$. One would like to learn a linear separating hyperplane classifier:

$$f(x) = sgn\ (w.x - b). \tag{1}$$

Additionally, this study wants this hyperplane to have the max.unraveling margin with respect to the two classes. Specifically, this study wants to locatethis hyperplane $H : y = w.x - b = 0$ and two hyperplanes parallel to it and with equal spaces to it,

$$\begin{aligned} &H_1 : y = w.x - b = +1, \\ &H_2 : y = w.x - b = -1 \end{aligned} \tag{2}$$

There should be no data points between $H_1$and$H_2$ , and the distance between $H_1$and $H_2$ should be as maximum as possible.

The coefficients vector w always be "normalized" so that the separating plane H and the corresponding $H_1$and$H_2$ in equation (2). Hence $H_1$will be $y = w.x - b = +1$and $H_2$ will be $y = w.x - b = -1$

Here the goal is to maximize the distance between $H_1$and $H_2$ . So that there will be some positive points on $H_1$and some negative points on $H_2$ . These points are called support vectors as they are the only ones participating in the definition of the separating hyper plane, and other points can be removed and / or moved around as long as they do not intersect the planes $H_1$and $H_2$

Recall that in 2-Dimentional, the distance from a point $(x_0, y_0)$to a line $Ax + By + C = 0$is

$$\frac{|Ax_0 + By_0 + C|}{\sqrt{A^2 + B^2}}. \tag{3}$$

In the same way, the distance of one point on $H_1$to $H : y = w.x - b = 0$is

$$\frac{|w.x - b|}{\|w\|} = \frac{1}{\|w\|}, \tag{4}$$

Andobviously the distance between $H_1$ and $H_2$ is

$$\frac{2}{\|w\|}. \tag{5}$$

So, to maximize the distance, one should minimize $\|w\| = w^T w$ and there should be no data points between $H_1$ and $H_2$:

$$\begin{aligned} &w.x - b \geq +1, \quad \textit{for positive examples}\, y_i = +1, \\ &w.x - b \leq -1, \quad \textit{for negative examples}\, y_i = -1. \end{aligned} \tag{6}$$

These two conditions can be united into

$$y_i\left(w.x - b\right) \geq 1 \tag{7}$$

So this problem can be written as

$$\min_{w,b} \frac{1}{2} w^T w \text{ subject to } y_i\left(w.x - b\right) \geq 1. \tag{8}$$

which is a convex and quadratic programming problem (in w,b), in a convex set.

By taking the Lagrange multipliers $\alpha_1, \alpha_2, ..., \alpha_N \geq 0$, the following Lagrangian can be obtained:

$$L(w,b,\alpha) \equiv \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i y_i \left(w.x_i - b\right) + \sum_{i=1}^{N} \alpha_i \tag{9}$$

### 2.1.2 The Dual Problem:

Maximize $L(w,b,\alpha)$ with respect to $\alpha$, subject to the restrictions that the gradient of $L(w,b,\alpha)$ with respect to w and b vanish:

$$\frac{\partial L}{\partial w} = 0 \tag{10}$$

$$\frac{\partial L}{\partial b} = 0 \tag{11}$$

And that

$$\alpha \geq 0$$

From equations 10 and 11, we have

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \, , \tag{12}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{13}$$

Substitute them into $L(w,b,\alpha)$, we have

$$L_D \equiv \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i . x_j \, , \tag{14}$$

in which the primal variables are eliminated.

When we solve $\alpha_i$, we can get

$$w = \sum\nolimits_{i=1}^{N} \alpha_i y_i x_i \, , \tag{15}$$

and we can classify a new object x with

$$f(x) = \operatorname{sgn}(w.x+b) = \operatorname{sgn}\left(\left(\sum_{i=1}^{N}\alpha_i y_i x_i\right).x+b\right) = \operatorname{sgn}\left(\left(\sum_{i=1}^{N}\alpha_i y_i (x_i.x)\right)+b\right) \qquad (16)$$

It is to be notedthat in the objective function and the solution, the training vectors $x_i$ occur only in the form of a dot product.

### 2.1.3 NON - LINEAR SVM:

The data points can be transformed to another high dimensional space so that the data points will be linearly separable. Let the transformation be $\Phi(.)$. At the high dimensional space, it can be solved as

$$L_D \equiv \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \Phi(x_i).\Phi(x_j) \qquad (17)$$

Suppose, in addition, $\Phi(x_i).\Phi(x_j) = k(x_i, x_j)$.this means, the dot product in the high dimensional space is equal to kernel function of input space. Until thekernel function $k(x_i, x_j)$ is equal to the dot product of the high dimensional space, one need not be explicit about the transformation $\Phi(.)$. There are various kernel functions that may be utilizedin this manner, for instance, the radial basis function, which is also known as Gaussian kernel

$$K(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma^2} \qquad (18)$$

The Mercer’s condition may be used to determine if a function can be used as a kernel function:

There exists a mapping $\Phi$ and an expansion

$$K(x, y) = \sum_i \Phi(x)_i \Phi(y)_i , \qquad (19)$$

if and only if, for any g(x) such that

$$\int g(x)^2 dx \text{ is finite.}$$

$$\text{Then } \int K(x,y)g(x)\,g(y)\,dx\,dy \geq 0 \qquad (20)$$

### 2.1.4 IMPERFECT SEPARATION:

The next step to extend SVM is to allow for imperfect separation or noise. That is, it is not firmly imposed that there may be no data points in between $H_1$ and $H_2$, but certainly reprimand the data points that cross the boundaries. The penalty C will be finite.

now non negative slack variables $\xi_i \geq 0$, be introduced so that

$$\begin{aligned} w.x - b &\geq +1 - \xi_i, \quad for\ \ y_i = +1, \\ w.x - b &\leq -1 + \xi_i, \quad for\ \ y_i = -1. \end{aligned} \qquad (21)$$

$$\xi_i \geq 0, \forall i$$

and add it to the objective function, which is a penalizing term:

$$\min_{w,b,\xi} imize \frac{1}{2} w^T w + C\left(\sum_i \xi_i\right)^m \qquad (22)$$

where m is set to 1 in general, which gives

$$\min_{w,b,\xi_i} imize \frac{1}{2} w^T w + C\left(\sum_{i=1}^{N} \xi_i\right)^m \qquad (23)$$

subject to

$$\begin{aligned} y_i\left(w^T x_i - b\right) + \xi_i - 1 \geq 0, \quad & 1 \leq i \leq N \\ \xi_i \geq 0, \quad & 1 \leq i \leq N \end{aligned} \tag{24}$$

Introducing Lagrange multipliers $\alpha, \beta,$ the Lagrangian is

$$L(w,b,\xi_i;\alpha,\beta) = \frac{1}{2} w^T w + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i \left[y_i\left(w^T x_i - b\right) + \xi_i - 1\right] - \sum_{i=1}^{N} \mu_i \xi_i$$

$$= \frac{1}{2} w^T w + \sum_{i=1}^{N} \left(C - \alpha_i - \mu_i\right)\xi_i - \left(\sum_{i=1}^{N} \alpha_i y_i x_i^T\right) - \left(\sum_{i=1}^{N} \alpha_i y_i\right) b + \sum_{i=1}^{N} \alpha_i \tag{25}$$

Neither the $\xi_i$ 's , nor their Lagrange multipliers, appear in the dual problem:

$$\max_{\alpha} imize\, L_D \equiv \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i . x_j$$

$$\text{Subject to } 0 \leq \alpha_i \leq C\,,\ \sum_i \alpha_i y_i = 0\,. \tag{26}$$

Now the $\alpha_i$ is bounded above by C instead of $\infty$. This is the only difference from perfectly separating case

The solution isyet again given by $w = \sum_{i=1}^{N} \alpha_i y_i x_i$ (27)

In order to train the Support Vector Machine (SVM), one may look through the feasible region of the dual problem and thus maximize the objective function. The optimal solution can be checked using KKT conditions.

### 2.1.5 THE KKT CONDITIONS:

The Karush–Kuhn–Tucker, also called as the Kuhn–Tucker conditions (KKT) optimality conditions of the primal problem are, the gradient of $L(w,b,\alpha,\beta)$ with respect to $w,b,\xi$ vanishes, and that for $1 \le i \le N$,

$$\alpha_i\left(y_i\left(w^T x_i - b\right) + \xi_i - 1\right) = 0 \qquad (28)$$

$$\mu_i \xi_i = 0 \qquad (29)$$

Now there are three cases left to consideron the basis of $\alpha_i$ value,

1. If $\alpha_i = 0$, then $\mu_i = C - \alpha_i = C > 0$. From equation (29), $\xi_i = 0$.
   So we have $y_i\left(w^T x_i - b\right) - 1 \ge 0$

2. If $0 < \alpha_i < C$, from equation (28), we have $y_i\left(w^T x_i - b\right) + \xi_i - 1 \ge 0$ (30)

   Note that $\mu_i = C - \alpha_i = C > 0$, so $\xi_i = 0$.

   Then (5) becomes $y_i\left(w^T x_i - b\right) - 1 = 0$

3. If $\alpha_i = C$, then from equation (28), we have $y_i\left(w^T x_i - b\right) + \xi_i - 1 = 0$ (31)

   Note that $\mu_i = C - \alpha_i = 0$, we have $\xi_i \ge 0$.

   So $y_i\left(w^T x_i - b\right) - 1 \le 0$

   The quantity $y_i\left(w^T x_i - b\right) - 1$can be computed as

   $$R_i = y_i\left(w^T x_i - b\right) - y_i^{\,2} = y_i\left(w^T x_i - b - y_i\right) = y_i E_i$$

   where $E_i = w^T x_i - b - y_i = u_i - y_i$ is the prediction error.

   Finally it can be concluded that the KKT condition implies

$$\alpha_i = 0 \Rightarrow R_i \geq 0$$
$$0 < \alpha_i < C \Rightarrow R_i \approx 0$$
$$\alpha_i = C \Rightarrow R_i \leq 0$$

The KKT condition is violated in the following two cases,

1. $\alpha_i < C \ \ and\ R_i < 0$
2. $\alpha_i > 0 \ \ and\ R_i > 0$

### 2.1.6 CHECKING KKT CONDITIONS WITHOUT UING THRESHOLD b:

Check the KKT condition without using threshold b, because the dual problem cannot solve for the threshold b directly.

The quantity $y_i\left(w^T x_i - b\right) - 1$(which should be $\geq 0$ *for all i* when the KKT condition is fulfilled) becomes as

$$y_i\left(w^T x_i - b\right) - 1 = y_i\left(w^T x_i - b\right) - y_i^{\ 2}$$
$$= y_i\left(w^T x_i - y_i - b\right)$$
$$= y_i\left(F_i - b\right)$$

where $F_i = w^T x_i - y_i$

Note for $E_i = F_i - b$ we have $E_i - E_j = F_i - F_j$ (32)

This notation is useful because the KKT conditions

$$\alpha_i = 0 \Rightarrow y_i\left(F_i - b\right) \geq 0$$
$$0 < \alpha_i < C \Rightarrow y_i\left(F_i - b\right) \approx 0$$
$$\alpha_i = C \Rightarrow y_i\left(F_i - b\right) \leq 0 \qquad (33)$$

Can be written as

$$i \in I_0 \ \ U \ \ I_1 \ \ U \ \ I_2 \Rightarrow F_i \geq b$$

$$i \in I_0 \ U \ I_3 \ U \ I_4 \Rightarrow F_i \leq b \qquad (34)$$

Where

$$I_0 \equiv \{i : 0 < \alpha_i < C\}$$
$$I_1 \equiv \{i : y_i = +1, < \alpha_i = 0\}$$
$$I_2 \equiv \{i : y_i = -1, < \alpha_i = C\}$$
$$I_3 \equiv \{i : y_i = +1, < \alpha_i = C\}$$
$$I_4 \equiv \{i : y_i = -1, < \alpha_i = 0\}$$

So that $\forall i \in I_0 \ U \ I_1 \ U \ I_2$ and $\forall j \in I_0 \ U \ I_3 \ U \ I_4$, we should have $F_i \geq F_j$, if KKT condition is satisfied.

In order to check whether this condition holds or not, define

$$b_{up} = \min\{F_i : i \in I_0 \ U \ I_1 \ U \ I_2\}$$
$$b_{low} = \min\{F_i : i \in I_0 \ U \ I_3 \ U \ I_4\} \qquad (35)$$

The KKT condition implies $b_{up} \geq b_{low}$, and similarly $\forall i \in I_0 \ U \ I_1 \ U \ I_2, F_i \geq b_{low}$ and $\forall i \in I_0 \ U \ I_3 \ U \ I_4, F_i \leq b_{up}$

Finally the comparisons without using the threshold b are got.

## 3. Feature Selection & It’s importance in classification problem:

One of the main problems of the NIDS is the enormous amount of data collected from the network that needs to be analyzed and tested for any potential intrusion [9], [10]. Therefore, in order to deal with this huge amount of data, it requires a system that can identify which features in the given dataset are useful or relevant, which features are useless or irrelevant. The system should be capable of identifying new threats quickly when such threats are introduced. The task of selecting the subset of relevant features from the total set of relevant and irrelevant feature is called 'Feature Selection'

The area of feature selection has attained huge attention by the researchers in machine learning over the last decade [11]. It can be perceived that from the literature, in so many machine learning and pattern recognition applications, the range of features are growing exponentially. It is quite obvious to mention that such huge amount of features contain many irrelevant features that eventually show effect on application performance. On these grounds researchers have been exploring various techniques in order to remove or at least reduce these irrelevant features to protect the system from negative effect.[12]. At the same time the features that are relevant and useful in the given task must be remained.

The main objective of the feature selection is to find out and select the optimal or best subset of features from the set of features. Even with this subset of features the classifier should be able to provide good prediction results [12], [13]. Based on the predefined constraints or criteria, Feature selection depends on a good search algorithm to find one or more usefulsubsets of features. This process can be presentedas follows:

Suppose if $F = \{F_1, F_2, ..., F_n\}$ be the entire set of features; then $S = \{F_{\tau(1)}, F_{\tau(2)}, ..., F_{\tau(m)}\}$,($S$⊆F) is a subset of features that is selected from the entire set, where $m<n$. The main aim of this subset selection is to select some features, which should form the most useful and informative subset ($S_{optimal} \subseteq F$) that could able to represent the actual data according to some criterion [14].

However, the original set of features may contain some irrelevant features. According to John, Kohavi, & Pfleger [15], a feature $F_i$ is considered relevant if and only if $p(Y = y / F_i = f_i) \neq p(Y = y)$ where $Y$ is the label, or output.

According to this definition, feature $F_i$ is relevant if its value can change the prediction for $Y$. To make it simple, '$Y$ is conditionally dependent of $F_i$'

The task of feature selection is very challenging. This task includes the selection of only useful features that are relevant to the prediction task [16]. If the total number of features are N then the number of feature subsets will be $2^N$. Once this 'N' gets larger, then exponential number of possible subsets will be generated. In such case comparing all these feature subsets to find out the best subset is the real challenge, especially when N>20. So it is very essential to find a good subset of features. As mentioned already, the feature election is made according to the feature relevance based on some evaluation criterion [6].

Actually, each feature has two possibilities: either IN (Selected) or OUT (Not selected). Therefore, the feature selection algorithm should select the best subset (IN) from the $2^N$ possible feature subsets. The ultimate goal is that this selected feature subset should give the best classification performance. A very important point at this juncture is, if the number features gets more and more, the problem becomes Non deterministic polynomial time hard (NP-Hard problem)

For any classifier, the classification task becomes harder if number of features is more. When the number of feature is more, there is every possibility of the presence of irrelevant features. These irrelevant features lead to the scenario of overfitting and confuse the classifier during the classification process. Overfitting is a phenomenon where the classifier predicts well on the training dataset and gives poor prediction on the testing data [17].

This typically occurs when there are many parameters in the model that the system is able to learn the data, rather than the trends in the data. This is especially common when there are many parameters in the system compared to the numbers of training examples or instances. Indeed, the number of parameters in the classifier is typically a function of the number of features.

There are many advantages of feature selection, one is that it helps in getting more understandingon the useful features and provides more insight on what the classifier does [18] ; [19]; [20]. Feature selection also helps in avoiding dimensionality problem that might arise due to too many features [21]; [12]. This dimensionality problem (having too many features) creates a phenomenon that extremely degrades the performance of various algorithms such as: classification, clustering, searching, and other in major operations used by applications of data mining applied on high dimensional data [22].

Apart from protecting from this huge problem of dimensionality that raises due to too many features in the dataset, feature selection is also very helpful in reducing the computational complexity and also helps in reducing the unnecessary cost of learning with those irrelevant features. It also provides good insight to understand the nature of the given problem [20]

Therefore it may be concluded that feature selection is one of the most important steps in any classification algorithms. It is observed that some feature selection methods are still failing with the overfitting phenomena and failing to give accurate results, which is as good as classification without feature selection.

At this juncture, it is very important to make a mention here that there is every possibility of poor classification results even after the feature selection. Sometimes the selected subset of features produces poor performance than the performance that is obtained without any feature selection. This is possible, especially in the case where all the features are useful and required for the task. Classifier also gives poor result in the case of selection of irrelevant feature subset to given problem. So selection of feature selection algorithm is very crucial.

In the present work two classification algorithms (RFE & RFA) were implemented to get relieved from the aforementioned problems. Implementation and results obtained through these algorithms are presented in the further sections.

## 3.1 Connection between Intrusion Detection and Feature Selection

It is very important to understand the useful connection that was identified by the researchers between Feature Selection and Intrusion Detection. One side, while the 1dimensional problem is converted into data set and on the other side machine learning algorithms are used with the intrusion detection problem, researchers established a useful connection between Feature Selection and Intrusion Detection.

Therefore, Intrusion Detection Problem is considered as a classification problem, which should be able to give a clear distinction between a normal connections and intrusions [23]. The performance of the classification process depends heavily on the features selection process. As already described in the previous section, the goal of feature selection algorithm is to remove the redundant and irrelevant features and consider only the useful and relevant features.

The classification algorithms gives very poor results if there are too many irrelevant features because the classifier will get confused with those many unnecessary features and starts giving wrong predictions. So it is very important to employ a good feature selection algorithm to improve the classification performance. That is the reason feature selection has become a part in NID systems especially while using machine learning algorithms for classification task.

A good feature selection process improves the Detection Rate (DR) and Accuracy (ACC) and decreases the False Alarm Rate (FAR). On the other side if there is no proper feature selection before classification, False Alarm Rate will be increased and Detection Rate and Accuracy will obviously get down. Hence the goal of feature selection with NIDS is to improve the Detection Accuracy by eliminating the irrelevant and repeated features (redundant features) from the actual data. As a result of applying feature selection on the entire dataset only a few set of features that consists only a subset of actual original features will be picked [24]. This is the reason feature selection has become an important pre processing step in many machine learning applications. As stated before, a data set that comprises of many features and few examples might lead to overfitting.

Usually, tracing new attacks on the Internet may reveal only few instances / examples. However, from those few examples, many features can be pulled out from the network traffic. Such attacks are certainly devastating to the networks and to the underlying systems. Therefore, it has become a necessity to find a solution for this problem.

In addition to this data set that was collected might contain some interdependent features that look no good but work very well when they are combined together. Therefore, the motivation of this work is to study few feature selection methods that can detect both independent and interdependent features out of a large set of features. These potential features are useful in intrusion detection to detect any novel threats that might compromise the Internet and its users because of the availability of limited number of examples. In other words the core classifier of intrusion detection need not deal with the entire set of features rather the classifier deals only with the subset of the total feature set. Application of feature selection with IDS can be done on the benchmark dataset or feature selection can be done from data that is on the network traffic. In the present work authors used the benchmark dataset ISCX 2012.

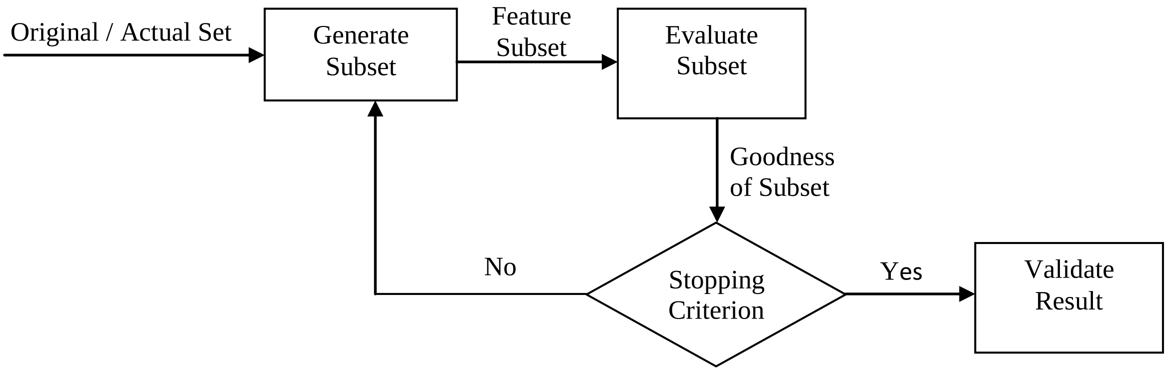


Figure 3: Four key-steps of feature selection

## 3.2 A New Trigram Technique - Handling Long Payload Features

In the previous sections, the need and importance of feature selection are elaborated. In the feature selection process, one of the major challenges is ‘handling long payload features’. Many of the machine learning algorithms cannot process such a long payload features. Because of this challenge, even some NID Systems are avoiding long payload features completely.

Network Intrusion Detection Systems operate on the real time network traffic, where having long payload features is a common phenomenon. Along with the length of the payload, these payloads are of different data types. This makes any machine learning classification algorithms highly difficult to deal with such long payload features

Due to the inability of handling these long payloads, avoiding them completely is not a good choice because many of the times, these long payloads carry important and sensitive information. Hence there should be a solution to this problem of handling long payload features.

Authors in [1] proposed a Bigram technique as a solution to this problem of dealing with long payload features during feature selection. Actually in the literature one might find the application of Bigram technique as an established technique in Deep Packet Inspection (DIP). This technique is studied over the decades [28]. The application of Bigram Technique in this context is proved to be useful as per the results presented in [1] but it certainly increases the computational overhead because of too many features that are generated.

However, in this work, a Trigram technique for the purpose of handling these long payload features is presented. In the literature, there is no work available on the usage of Trigram technique; hence the authors have made an attempt to apply Trigram techniques in the task of handling long payload features during feature selection process. Unlike taking two adjacent words in Bigram from a feature, in Trigram three adjacent words are taken. The following section demonstrates the generation of Trigrams from a feature.

### 3.2.1 Generation of Trigrams:

Here, the three payload features ABC2ZvL, FueXljc, 2ZveXl are taken as example. Just for explanation these features are taken, readers are informed not to get confused with the length as these payloads as they are not long payloads. The resulted Trigrams look as follows:

ABC | BC2 | C2Z | 2Zv | ZvL | Fue | ueX | eXl | Xlj | ljc | Zve | veX . The Trigrams that are repeated are excluded from this list. Table 1 gives better picture of the resulted Trigram features.

**Table 1: Trigram representation for the three payload features in the example**

| **Original payload** | ABC | BC2 | C2Z | 2Zv | ZvL | Fue | ueX | eXl | Xlj | ljc | Zve | veX |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ABC2ZvL | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| FueXljc | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 2ZveXl | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |

Table 1 shows the encoded Trigram features from the given three payloads. The header of the table is the standard feature vector for the total payload features in the given example. The first row represents the first payload feature; second row corresponds to the second payload feature and so on.

The very first step while handling these long payload features during feature selection process is construction of dictionary. So a dictionary is built with all the Trigrams in it, which generates a feature vector. The construction of Trigram Dictionary is elaborated with examples and various results in [31] by the authors of this work.

The feature selection process while handling long payload feature is depicted in figure 4. First the long payload features are given as input then a dictionary is built. Using the dictionary Feature Vectors are built. The system selects these feature vectors that are created out of the long payload features.

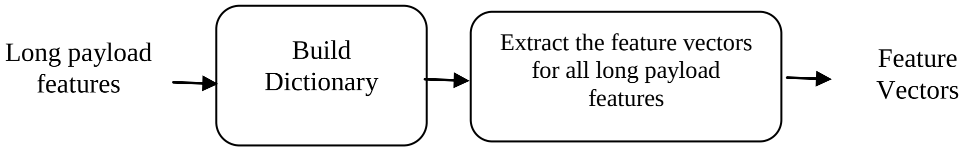


Figure 4: Feature extraction process for ISCX data set using Trigram technique

As mentioned earlier, the algorithm for Trigram Dictionary construction can be found in [31]

Once the Trigram dictionary is built, that can be used by any feature selection algorithms even more efficiently. Feature selection algorithms perform well with Trigrams rather with long payload features. The algorithm of Feature Vector Extraction for these long Payload Features with Trigrams is presented in [31] by the authors. Now any machine learning classification algorithms can use the generated trigrams to effectively classify. In the present work Support Vector Machine is used for classification and various experiments are conducted. Some results are presented at the end of this chapter.

Even though some Java codes were written for these methodologies to experiment and in many other experiments of the present work WEKA GUI v3.8 has been used [29].

Once this preprocessing task is done, the data set should be prepared for feature selection. For this purpose a pre ranking of features using the well known Correlation Based Feature Selection (CFS) is done. CFS is a fast feature selection algorithm that has been widely used in the literature [30]. This pre ranking step is very important because even though the features are reduced, still one has to make sure that these reduced features are still relevant and useful.

By using this method, it is easy to rank the features to perform controlled experiments through manipulating relevant features and irrelevant features [30]. As it is one of the old procedures, the explanation of the same is omitted here. The detailed description and more results of comparisons of Trigram with Bigram Techniques were given by the authors in [31]

## 3.3 Analysis of Recursive Feature Elimination (RFE) & Recursive Feature Addition (RFA)

In the year 2002, Guyon, I., Weston, J., Barnhill [4] had presented a new approach of gene selection with Support Vector Machine using a new feature Selection Method, called Recursive Feature Elimination (RFE). In the year 2007, Chen & Jeong [3] identified a tendency of RFE as it removes some interdependent features. That means RFE has the tendency of removing some feature that might look useless individually but they work effectively with the combination of other features.

This gave the motivation to one of the recent works of Tarfa Hamed, Rozita Dara, Stefan C. Kremer to propose and experiment a new feature selection, called ‘Recursive Feature Addition [1].

For analyzing the performance and to investigate the effectiveness of RFE and RFA, some synthetic data sets were designed in such a way that some features work well in combination and applied such datasets on RFE to solve it. The pliability of RFE has been checked to noisy and irrelevant features and has been tested its ability to find interdependent features using a majority problem. But it is observed that RFE algorithm fails to identify the interdependent features. RFE algorithm focuses more on selecting individual features, whereas RFA performed little better than RFE in this scenario of selecting interdependent features. The two algorithms, RFE and RFA are presented below.

**Algorithm 1**: Recursive Feature Elimination Algorithm [4]

```
1.  Give Input Training set $T_s$ , Train the model on the $T_s$ using all predictors
2.  Take the actual or Original Feature set F
3.  Calculate the model performance $P_m$ on $T_s$
4.  Calculate rankings to assign variable importance
5.  Let $S_i$ be a subset of feature set 'F', for each subset size $S_i$,
6.  While  i <= S do
7.  Keep the $S_i$ most important variables
8.  Pre-process the data [Optional]
9.  Train the model on the training set using  $S_i$  predictors
10. Calculate model performance
11. Recalculate the rankings for each predictor [Optional]
12. Endwhile
13. Calculate the performance profile over the $S_i$
14. Determine the appropriate number of predictors
15. Use the model corresponding to the optimal $S_i$
```

**Algorithm 2**: Recursive Feature Addition (RFA) [1]

```
1.  Give input  $D \leftarrow Dataset$,  $F \leftarrow actual\ feature\ set$ ,  $N \leftarrow No.of\ \ features$
2.  $S \leftarrow Empty$ ('S' consists the ranked feature set)
3.  $i \leftarrow 1$
4.  while  $i \leq N$  do
4.  Train Support Vector Machine's classifier
5.  Get the resulted vector,  $\alpha \leftarrow$ alpha vector
6.  Get the resulted support vectors  $X \leftarrow$ support vectors
8.  Calculate the ranking coefficients for the remaining features
9.  Add the feature '$f$' that has the maximum ranking coefficient to the set S,  $S \leftarrow S + f$
10. Remove feature f from actual features set F,  $F \leftarrow F - f$
11. $i \leftarrow i + 1$
12. Endwhile
13. Output:  Get the Ranked features set S.
```

In the present work, the above two feature selection algorithms are empirically studied. The benchmark ISCX 2012 datasets are used for conducting various experiments in this study. Both the algorithms are exposed to various synthetic and real datasets with extreme conditions to test the efficacy of both the algorithms. SVM is used for the classification purpose for both the feature selection algorithms. Below section presents the SVM as classifier.

## 4. Using SVM as a Classifier in the Present work:

In section 3, the theoretical development and mathematical formulation have been presented. This section deals with the application of SVM in the present work. The mathematical notations in this section will slightly vary with the notations in the previous section, the intentionis to present the concept with more clarity in its application.

In the present work, the Support Vector Machine (SVM) has been extensively used in solving machine learning problems due to its accurate classification results. This classifier finds the maximum margin between training examples and the decision boundary [17]. The decision boundary of a SVM classifier is illustrated in Figure 2.

SVM is used by various researchers over the decades in regression, classification and other machine learning tasks [5]. The winner in ICML 2013 in the facial expression challenge used the primal objective of a SVM as the loss function for training [25]. In the following sections, Support Vector Machine (SVM) is presented in the context of present work.

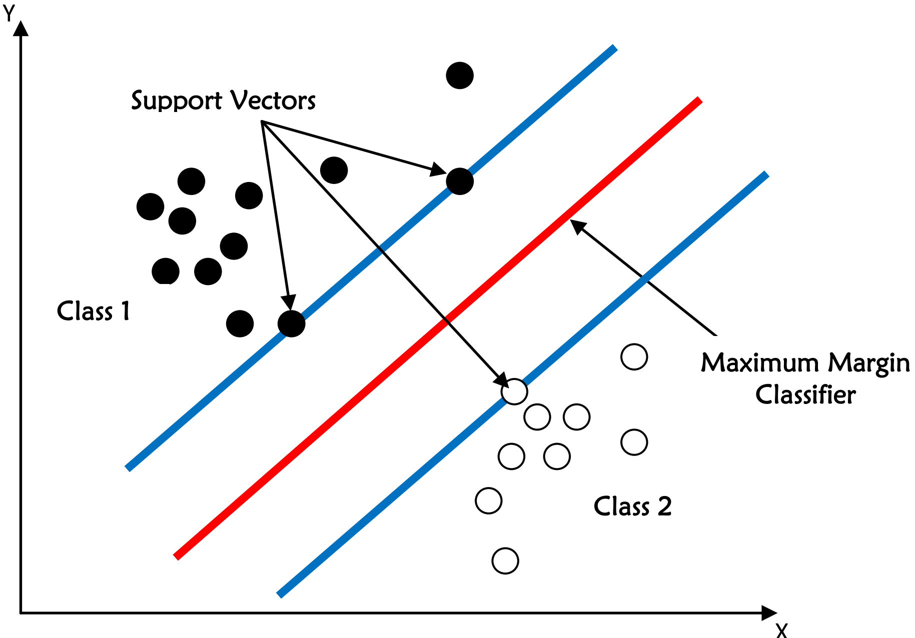


Figure 4: Decision boundary and margin of SVM classifier

Given training examples $x_i \in R^n, i = 1,...,l$ that belong to two classes, and a class label vector $y \in R^l$ such that $y_i \in \{1,-1\}$, according to Boser et al. [17], Support Vector Classification (C_SVC) can solve this optimization problem:

$$\underset{w,b,\xi}{\min}\left(\frac{1}{2}w^T w + C\sum_{i=1}^{l}\xi_i\right) \qquad (36)$$

Subject to

$$y_i\left(w^T\,\Phi(x_i) + b\right) \geq 1 - \xi_i \;,\; \xi_i \geq 0, i = 1,...,l$$

where ($x_i$) projects $x_i$into a higher dimensional space, C > 0 is used as a regularization parameter. Since the vector variable w can be of high dimensionality, the dual problem is solved as follows:

$$\underset{\alpha}{\min}\left(\frac{1}{2}\alpha^T Q\alpha - e^T\alpha\right) \qquad (37)$$

Subject to

$$y^T\alpha = 0,\; 0 \leq \alpha_i \leq C, i = 1,...,l$$

Where $e = [1,\, ...,1]^T$ is a vector of all ones and Q is an $l\;X\;l$ positive matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$, and $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$is the kernel function.

Once the above problem is solved, the primal –the dual relationship may be used to find the optimal $\omega$ which satisfies:

$$\omega = \sum_{i=1}^{l} y_i\,\alpha_i\,\phi(x_i) \qquad (38)$$

Then the decision function will be:

$$\operatorname{sgn}\left(\omega^T(x)+b\right)=\operatorname{sgn}\left(\sum_{i=1}^{l} y_i\,\alpha_i\,K(x_i,x)+b\right) \tag{39}$$

The SVM classifier is trained according to below procedure:

5. Procedure SVM_train (Training examples {$x_1$, $x_2$,…, $x_{l\}}$ , class labels {$y_1$, $y_2$,…,$y_l$})
6. Minimize over $\alpha_k$ :

$$J=\left(\frac{1}{2}\right)\sum_{hk} y_h y_k \alpha_h \alpha_k \left(x_h.y_k+\lambda\delta_{hk}\right)-\sum_{k}\alpha_k \tag{40}$$

7. Outputs: Parameters $\alpha_k$

Where $x_h.y_k$ denotes the scalar product, $y_k$ corresponds tothe class label as a binary value of +1 (or) -1, the summations are applied on all training examples $x_k$ which are feature vectors with n-dimensions each, $\delta_{hk}$ is the Kronecker symbol ($\delta_{hk}$ = 1 if h = k otherwise $\delta_{hk}$ = 0), and $\lambda$ and C are positive constants. In general, the decision function of an input vector x using SVMs is:

$$D(x) = w.x + b \tag{41}$$

with $\omega=\sum_k \alpha_k y_k x_k$ and $b=\left(y_k-\omega.x_k\right)$ where $\omega$ is the weight vector which consists of a linear set of training examples. The training examples that have non-zero weights are considered as support vectors [17]. In addition, using an SVM classifier in the process of feature selection is a good idea due to the great classification performance of this classifier and due to the ubiquitous use of SVMs in a wide range of application domains and problems [25]; [26].

## 4.1 ISCX 2012 data set:

The Information Security Centre of Excellence (ISCX) at the University of New Brunswick generated the ISCX 2012 data set in 2012 [2]). The data set of ISCX involves actual traces that were examined to generate profiles for agents that produce real traffic for HTTP, FTP, SMTP, SSH, POP3 and IMAP.

This Dataset consists of many features such as number of bytes, which are send or received. Apart from these relevant features, the data set, which was generated includes various feature like full packet payloads, etc., For about seven days this complete dataset has been recorded. There are twenty features and a total of one million network trace packets in this dataset. Every data example is labeled as either of the two classes; attack or normal.

Due to its labeled connections, multiple attack scenarios and realistic traffic,the ISCX data set has obtained the security community's attention and considered as a benchmark data set for the research purpose of intrusion detection. The ISCX data set is superior to other datasets that are openly available because these data sets were designed to defeatthe technical limitations of other data sets that were also created for Intrusion Detection purposes. The ISCX dataset was prepared through network traces by capturing contemporary genuineand intrusive network patterns and behaviors.

## 5. Composite Performance Evaluation Metric:

Researchers have been contributing various solutions towards addressing this problem. Network Intrusion Detection Systems (NIDS) is one of the better solutions against these cyber crimes and network attacks. At this juncture it is important to note that there should be an effective metric to evaluate the performance of these Network Intrusion Detection Systems.

In the literature there are number of primitive metrics are available, that include accuracy, F- Measure, Detection Rate and so on. But it is unfortunate to note that in spite of decades of research in the area of NIDS, there is no benchmark performance evaluation metric is available in the literature. While researchers are coming up with new approaches of classification, it is important to have an established metric to evaluate the performance classification to determine the efficiency of NID Systems.

In this present research investigation, a new Composite Performance Metric, which is called 'DAFAR' is proposed by the authors to evaluate the performance of different classifiers in Network Intrusion Detection Systems.

Recently A new combined metric has been proposed in [1], which is called 'combined'. The formula has been developed by incorporating three metrics together (Accuracy, Detection rate, and FAR)

$$Combined = \left( \frac{ACC + DR}{2} \right) - FAR \tag{42}$$

In this formula, equal weights to all the three metrics (accuracy, detection rate and false alarm rate) are given. The authors explained with different example scenarios in [1] thatthis formula directs to determine the performance when there are cases of equal accuracy and equal False Alarm Rate. This looks like a balanced way of measuring the performance of certain applications. But it might not work well with the other applications. The major loop hole with this combined metric is discussed in the further sections of this paper. It is preferable to have an evaluation metric / measure that works for if not for all but for the majority of the applications.

After many experiments on the benchmark data set and after observing the results under large number of scenarios, the composite performance metric is proposed as presented in equation (43). This proposed Metric is named as "**DAFAR**".

$$DAFAR = \left( \frac{ACC + (M_f * DR)}{2} \right) - \frac{FAR}{2} \tag{43}$$

'DAFAR', stands for D-Detection, A-Accuracy, and FAR-False Alarm Rate. $M_f$ is the multiplication factor. It can be noted from the equation (43) that more weight is given to Detection Rate over other two primitive metrics. It is quite simple to understand that in any NIDS, detecting an intrusion is the key factor because the very objective of the NIDS is to detect an intrusion. The proposed 'DAFAR' is observed to be very effective in evaluating the performance of various classifiers over different scenarios.

After many experiments on the benchmark datasets, the value of the multiplication factor $M_f$ is set to '2'. The proposed DAFAR approach gives better results when the value of $M_f$ is 2. The value of DAFAR is always a positive real number between 0 to 1.5 and in most of the cases, the value lies between 0 to 1 except in exceptional scenarios.

Experimental results proved that the proposed DAFAR Metric gives clear direction on the performance of Network Intrusion Detection System (NIDS) over other ID systems. Authors in [1] considered an example, where a data set of 200 instances are taken and split into 100 normal instances and 100 attack instances, then discussed three different scenarios that have the same accuracy of 50%.

The confusion matrixes are taken as below three scenarios:

**Scenario 1:**

$$Confusion\,matrix = \begin{bmatrix} 1 & 99 \\ 1 & 99 \end{bmatrix}$$

**Scenario 2:**

$$Confusion\,matrix = \begin{bmatrix} 50 & 50 \\ 50 & 50 \end{bmatrix}$$

**Scenario 3:**

$$Confusion\,matrix = \begin{bmatrix} 70 & 30 \\ 70 & 30 \end{bmatrix}$$

For all the three different confusion matrices that are considered, the primitive and combined performance metrics' values were calculated [1] as per the Equation (42) as shown in Table 2

**Table 2: Performance metrics for three different scenarios with equal accuracy**

| Scenario | Accuracy | DR | FAR | Combined |
|---|---|---|---|---|
| Scenario 1 | 0.5 | 0.99 | 0.99 | −0.245 |
| Scenario 2 | 0.5 | 0.5 | 0.5 | 0 |
| Scenario 3 | 0.5 | 0.3 | 0.3 | 0.1 |

Looking at the table, one may get confused to select which scenario is better based on accuracy because all the three have equal values. For this reason a combined metric was introduced by the authors in [1]. As per this metric, third scenario was selected. It is important to note that selecting a classifier whose Detection Rate (DR) is just 0.3 was selected as better, leaving the first scenario who Detection Rate (DR) is 0.99.

The proposed new composite metric as presented in equation (43) leads to a better direction in selection process of one classifier over the other. Which ever gets the highest DAFAR value, will be selected as better one. Table 3 describes the DAFAR calculation in the taken three scenarios.

**Table 3: Performance metrics for three different scenarios with DAFAR Performance Metric**

| Scenario | Accuracy | DR | FAR | Combined | DAFAR |
|---|---|---|---|---|---|
| Scenario 1 | 0.5 | 0.99 | 0.99 | −0.245 | 0.745 |
| Scenario 2 | 0.5 | 0.5 | 0.5 | 0 | 0.5 |
| Scenario 3 | 0.5 | 0.3 | 0.3 | 0.1 | 0.4 |

It is quite obvious to note that based on the proposed new composite metric, the first scenario is selected. A total of hundred scenarios are taken in the present research work and done the calculation with the primitive, combined and proposed new DAFAR metric to prove the efficacy of the proposed method. The detailed explanation with all the results of this work is presented by the authors in [32]. Another major adventage with DAFAR as performance metric is, it always gives a positive real value between 0 to 1.5, where in the most of the cases the value lies between 0.5 to 1 that established a strong correlation with other primitive metrics like accuracy and detection rate.

## 6. Experimental Results:

### Results of RFE &RFA application on ISCX data set:

Different experiments were carried out on two different feature selection (RFE & RFA) approaches with the same SVM classifier. In order to observe the effect of handling the payload features in improving the detection accuracy, a crucial experiment was conducted. First the data subsets were created in such a way that the subset consists less number of examples and more number of features. This is a difficult scenario, butthese experiments were carried outto test the overfitting effect. In both RFE & RFA application, the classifier underperformed in this scenario. But a slightly better performance is observed with RFA but with a narrow margin in the performance that is observed through empirical results, it is hard to judge one approach asbetter over the other. But with respect to the selection of interdependent feature selection, certainly RFA seems to be superior.Table4 presents the performance of SVM with RFE feature selection on ISCX 2012 datasets. Resulted performances are presented through different primitive metrics such as Accuracy (ACC), Specificity, F-Measure, Detection Rate (DR) and False Alarm Rate (FAR)

**Results of RFE application on ISCX 2012 Dataset:**

**Table 4: Performance metrics of SVM with RFE on the ISCX data sets**

| Data Set | Performance Metrics | | | | |
|---|---|---|---|---|---|
| | Accuracy | Specificity | F – measure | Detection Rate | False Alarm rate |
| 20 examples | 76.60% | 72.25% | 76.00% | 73.36% | 47.36% |
| 40 examples | 87.70% | 86.98% | 88.21% | 86.00% | 42.58% |
| 100 examples | 88.20% | 87.12% | 88.20% | 86.85% | 38.32% |
| 400 examples | 92.01% | 90.05% | 91.90% | 89.54% | 29.32% |
| 800 examples | 93.96% | 91.25% | 92.97% | 91.02% | 23.85% |
| 1200 examples | 95.66% | 93.01% | 95.01% | 94.42% | 20.09% |

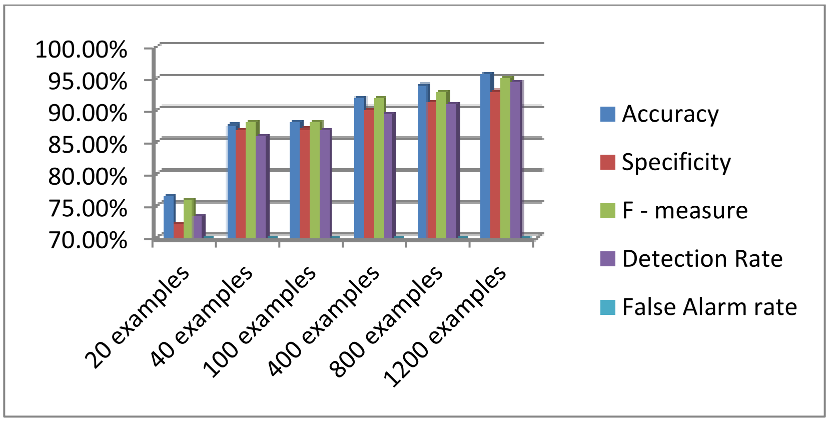
**Results of RFA application on ISCX 2012 Dataset:**

**Table 5: Performance metrics of SVM with RFA on the ISCX data sets**

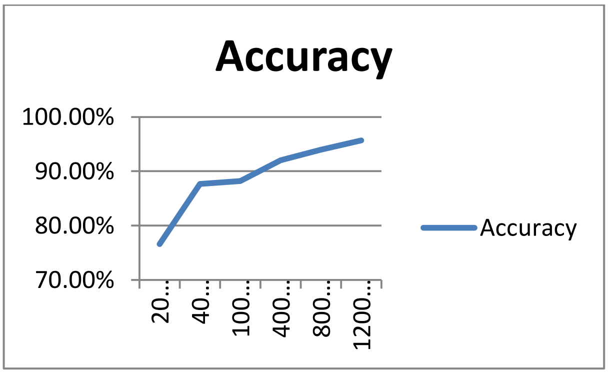
| Data Set | Performance Metrics | | | | |
|---|---|---|---|---|---|
| | Accuracy | Specificity | F – measure | Detection Rate | False Alarm rate |
| 20 examples | 77.60% | 74.32% | 77.00% | 75.82% | 48.54% |
| 40 examples | 88.60% | 85.65% | 88.60% | 86.12% | 43.65% |
| 100 examples | 88.80% | 86.12% | 88.70% | 88.00% | 39.00% |
| 400 examples | 92.90% | 90.41% | 92.90% | 91.02% | 30.23% |
| 800 examples | 94.85% | 91.98% | 93.92% | 92.00% | 24.26% |
| 1200 examples | 96.12% | 93.07% | 95.99% | 94.75% | 19.32% |

Table 5 presents the performance of SVM with RFA feature selection on ISCX 2012 datasets. It can be observed from the table that six datasets with different sizes are used in the experiments. It may be noted from the above two tables that performance is better when there aremore number of instances. When there are less examples, performances of SVM classification is degraded with both the feature selection (RFE & RFA) approaches. Still it seems the results look good because in the present section of experiments number of features have not increased nor decreased. The following graphs give even better picture on the performance of classification with the presented two approaches.

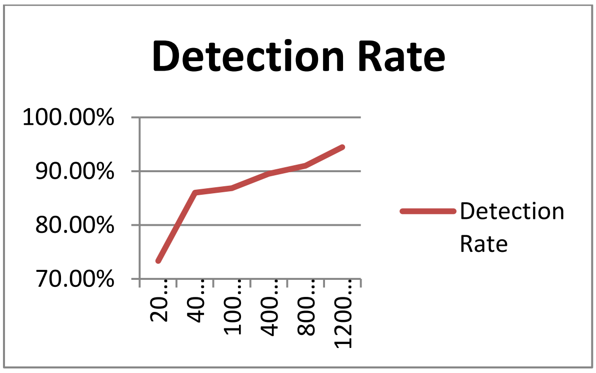
**Performance of SVM with RFE on ISCX 2012 Dataset:**



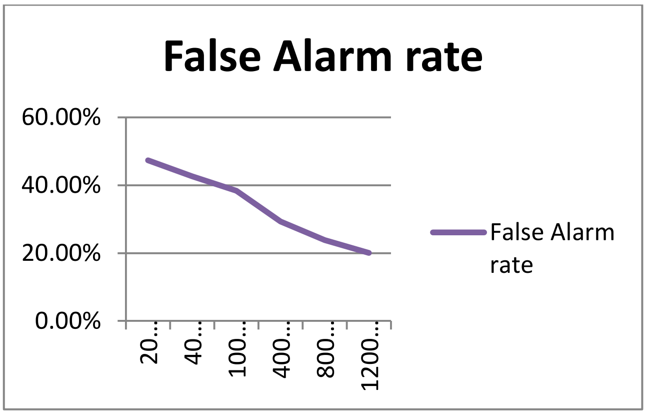
Graph 1: Performance metrics of SVM with RFE on the ISCX data sets



Graph 2: Accuracy of SVM with RFE on ISCX datasets

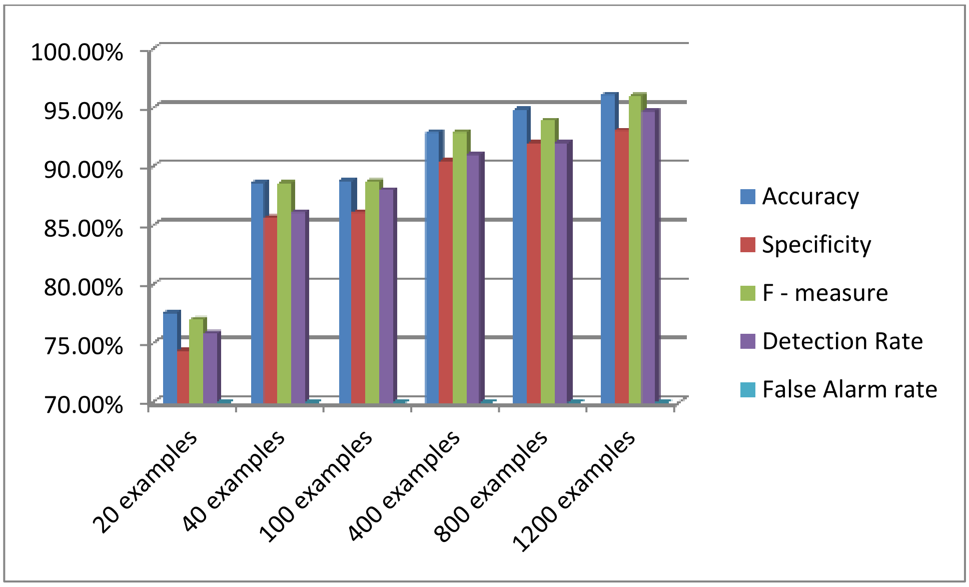


Graph 3: Accuracy of SVM with RFE on ISCX datasets

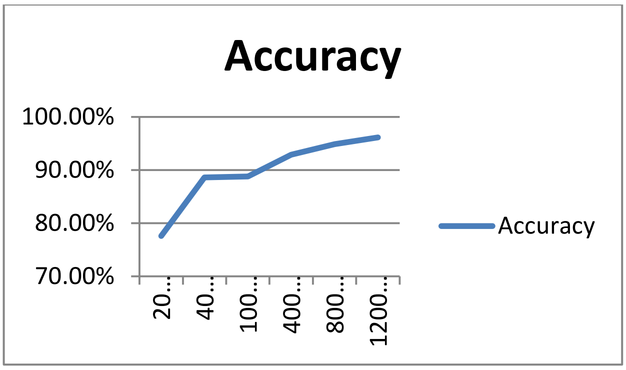


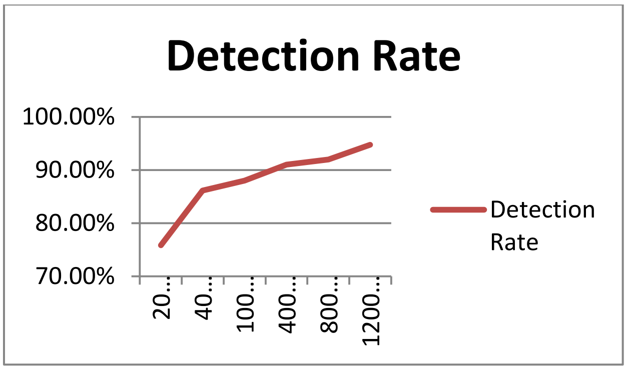
Graph 4: FAR of SVM with RFE on ISCX datasets

**Performance of SVM with RFA on ISCX2012 Dataset:**



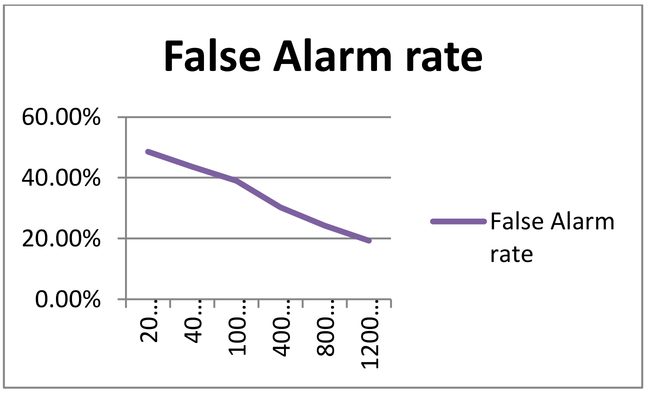
Graph 5: Performance metrics of SVM with RFA on the ISCX data sets





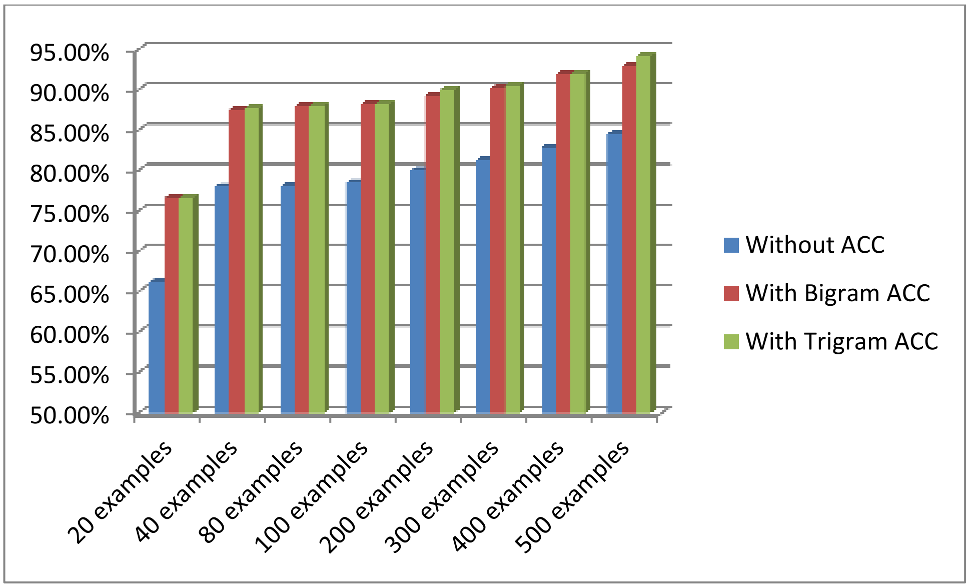
Graph 6: Accuracy of SVM with RFA on ISCX datasets

Graph 7: Accuracy of SVM with RFA on ISCX datasets

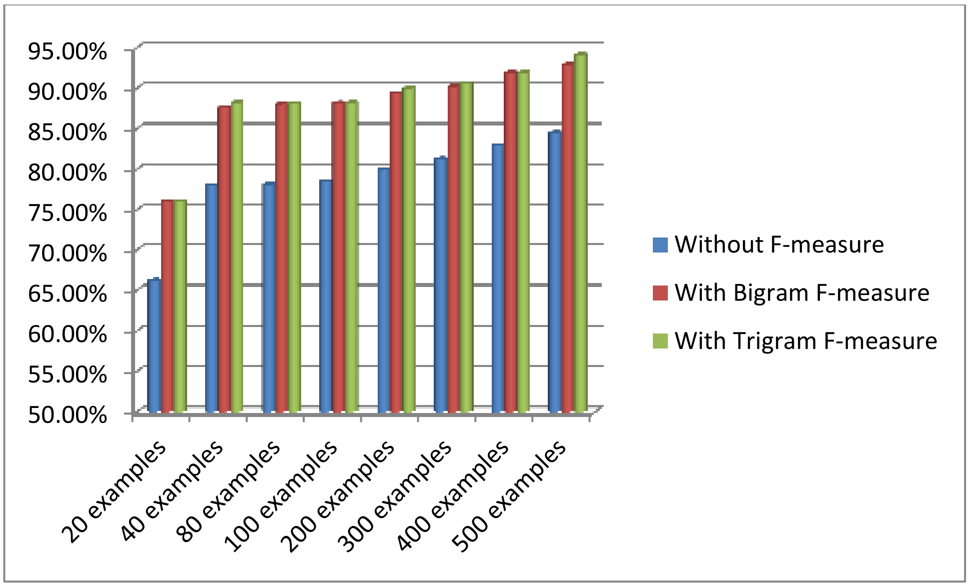


Graph 8: FAR of SVM with RFA on ISCX datasets

## Analysis between Trigram and Bigram Techniques - handling long payload features



Graph 9: Accuracy without encoding, with Bigram, with Trigram

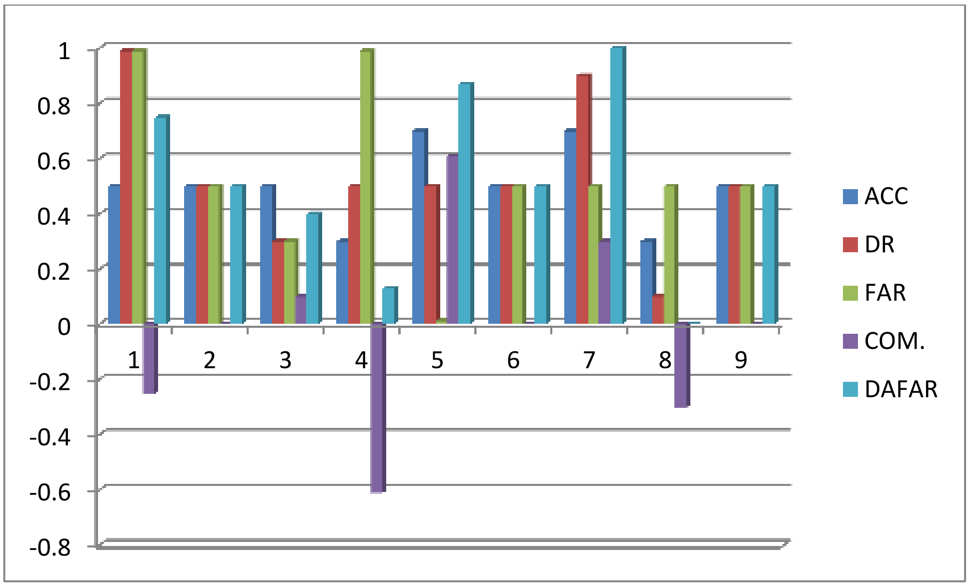


Graph 10: F-Measure without encoding, with Bigram, with Trigram

### An important observation:

When there are more number of similar string values in the payload string, Bigram slightly performs well, whereas in other cases Trigram performs better. After many number of experiments, it is observed that using Trigram scheme leads to better results in classification while reducing the computational overhead

**A composite Performance evaluation metric DAFAR and other metrics:**



Graph 11: DAFAR and other performance metrics on three scenarios

DAFAR with the multiplication factor $M_f$ of 2, always has its value a positive real numbers between 0 to 1 in most of the cases, in exceptional cases the value may go up to a maximum extent of 1.5.

## 7. Conclusion & Future Scope:

In this paper, a rigorous study and analysis of the basic concepts in Network Intrusion Detection Systems and fundamental concepts in feature selection are presented. The theoretical development and mathematical formulations were well explained in this paper. The Underlying mathematical bases for the classification algorithms are also presented. A new Trigram technique was designed to encode long payload features. Experiments were carried out with two major feature selection approaches, Recursive Feature Elimination (RFE) and the latest Recursive Feature addition (RFA) on dataset ISCX 2012 by using well established SVM classification algorithms. Comparisons are made through the empirical results. It is observed that Recursive Feature Elimination (RFE) is performing well while handling independent features and Recursive Feature Addition (RFA) is performing well while handling the interdependent features. In any NIDS, both independent and interdependent features are very important. In this context a new hybrid approach may be proposed by optimizing the advantages of both the algorithms, which is left as a future scope of this work. An intelligent system may also be developed to choose various methods / schemes dynamically based on the task and dataset.

**References:**

[1] TarfaHamed , Rozita Dara, Stefan C. Kremer, Network intrusion detection system based on recursive feature addition and bigram technique, computers & security 73 (2018) 137–155

[2] Shiravi A, Shiravi H, Tavallaee M, Ghorbani AA. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. ComputSecur 2012;31(3):357–74 doi. Available_from: http://dx.doi.org/10.1016/j.cose.2011.12.012,http://www.sciencedirect.com/science/article/ pii/S0167404811001672.

[3] Chen, X.-w., &Jeong, J. C. (2007). Enhanced recursive feature elimination. In Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on, (pp. 429{435).

[4] Guyon, I., Weston, J., Barnhill, S. et al., Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning (2002) 46: 389. https://doi.org/10.1023/A:1012487302797.

[5] Chang, C.-C., & Lin, C.-J. (2011). Libsvm: a library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2 (3), 27.

[6] Chandrashekar, G., &Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40 (1), 16 - 28. 40th-year commemorative issue. URL http://www.sciencedirect.com/science/article/pii/S0045790613003066

[7] Mathews Lee , (2017, July 13), Millions Of Verizon Customers Exposed By Third-Party Data Leak, Retrieved fromhttps://www.forbes.com/sites/leemathews/2017/07/13/millions-of-verizon-customers-exposed-by-third-party-leak/#35fbeb232836.

[8] Ponemon Institute LLC (2017) “2017 Cost of Data Breach Study Global Overview”, Benchmark research sponsored by IBM Security Independently conducted by Ponemon Institute LLC.

[9] Singh, S., &Silakari, S. (2009). A survey of cyber attack detection systems. International Journal of Computer Science and Network Security, 9 (5), 1 - 10.

[10] Kayacik, H. G., Zincir-Heywood, A. N., & Heywood, M. I. (2005). Selecting features for intrusion detection: A feature relevance analysis on kdd 99 intrusion detection datasets. In Proceedings of the Third Annual Conference on Privacy, Security and Trust (PST-2005).

[11] Shanab, A., Khoshgoftaar, T., & Wald, R. (2011). Impact of noise and data sampling on stability of feature selection. In Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on, vol. 1, (pp. 172{177).

[12] Chandrashekar, G., &Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40 (1), 16 - 28. 40th-year commemorative issue. URL http://www.sciencedirect.com/science/article/pii/S0045790613003066

[13] Moustakidis, S., &Theocharis, J. (2012). A fast svm-based wrapper feature selection method driven by a fuzzy complementary criterion. Pattern Analysis and Applications, 15 (4), 379{397.URL http://dx.doi.org/10.1007/s10044-012-0293-7

[14] Zeng, Z., Zhang, H., Zhang, R., & Yin, C. (2015).A novel feature selection method considering feature interaction. Pattern Recognition, 48 (8), 2656 { 2666. URL http://www.sciencedirect.com/science/article/pii/S0031320315000850

[15] John, G. H., Kohavi, R., & P eger, K. (1994). Irrelevant features and the subset selection problem. In MACHINE LEARNING: PROCEEDINGS OF THE ELEVENTH INTER-NATIOFINAL, (pp. 121{129). Morgan Kaufmann.

[16] Yang, Q., Salehi, E., & Gras, R. (2010). Articial Intelligence and Soft Computing: 10th International Conference, ICAISC 2010, Zakopane, Poland, June 13-17, 2010, Part I , chap. Using Feature Selection Approaches to Find the Dependent Features, (pp. 487{ 494). Berlin, Heidelberg: Springer Berlin Heidelberg. URL http://dx.doi.org/10.1007/978-3-642-13208-7_61

[17] Boser, B. E., Guyon, I. M., &Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory, (pp. 144{152). ACM.Canada, S. (2013). Canadian internet use

survey, 2012. Accessed: 2016-09-06. URL http://www.statcan.gc.ca/daily-quotidien/131126/dq131126d-eng.htm

[18] Moustakidis, S., &Theocharis, J. (2012). A fast svm-based wrapper feature selection method driven by a fuzzy complementary criterion. Pattern Analysis and Applications, 15 (4), 379-397.URL http://dx.doi.org/10.1007/s10044-012-0293-7

[19] Wang, S., Pedrycz, W., Zhu, Q., & Zhu, W. (2015). Subspace learning for unsupervised feature selection via matrix factorization. Pattern Recognition, 48 (1), 10 { 19. URL http://www.sciencedirect.com/science/article/pii/S0031320314002957

[20] Yao, J., Mao, Q., Goodison, S., Mai, V., & Sun, Y. (2015). Feature selection for unsupervised learning through local learning. Pattern Recognition Letters, 53 , 100 - 107. URL http://www.sciencedirect.com/science/article/pii/S0167865514003559

[21] Bolon-Canedo, V., Sanchez-Marono, N., & Alonso-Betanzos, A. (2011). Feature selection and classification in multiple class datasets: An application to fKDDg cup 99 dataset. Expert Systems with Applications, 38 (5), 5947 – 5957

[22] Bernecker, T., Houle, M. E., Kriegel, H.-P., Kr•oger, P., Renz, M., Schubert, E., &Zimek, A. (2011). Quality of Similarity Rankings in Time Series, (pp. 422{440). Berlin, Heidelberg:Springer Berlin Heidelberg. URL http://dx.doi.org/10.1007/978-3-642-22922-0_25

[23] UdayBabu P., V. R., Priya C. G. (2014). Survey on intrusion detection techniques using data-mining domain. International Journal of Engineering Research & Technology, 3 (4), 1067-1070.

[24] Mzila, P., & Dube, E. (2013). The effect of destination linked feature selection in real-time network intrusion detection. In ICIMP 2013 : 8th International Conference on Internet Monitoring and Protection. Think Mind 2013.

[25] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., &Bengio, Y. (2015). Challenges in representation learning: A report on three machine learning contests. Neural Networks, 64 , 59 - 63. Special Issue on Deep Learning of

Representations.URL http://www.sciencedirect.com/science/article/pii/S0893608014002159

[26] Chamasemani, F. F., & Singh, Y. P. (2011). Multi-class support vector machine (svm) classifiers { an application in hypothyroid detection and classification. In Proceedings of the 2011 Sixth International Conference on Bio-Inspired Computing: Theories and Ap-plications, BIC-TA '11, (pp. 351-356). Washington, DC, USA: IEEE Computer Society. URL http://dx.doi.org/10.1109/BIC-TA.2011.51

[27] Ebook: Support vector machine, Anusha Bhardwaj, Indian Agricultural Statisticss Research Institute, New Delhi, pp 171-179, Support vector machine for binary classification is based on Burges (1998). http://www.iasri.res.in/ebook/fet/Chap%2016_SUPPORT%20VECTOR%20MACHINES.pdf

[28] NiccolòCascarano, Luigi Ciminiera, and FulvioRisso. 2010. Improving cost and accuracy of DPI traffic classifiers. In Proceedings of the 2010 ACM Symposium on Applied Computing (SAC '10). ACM, New York, NY, USA,641-646.DOI=http://dx.doi.org/10.1145/1774088.1774223

[29] Open Source, freely available and downloadable from: https://www.cs.waikato.ac.nz/ml/weka/downloading.html

[30] Hall, M. A. (1999). Correlation-based feature selection for machine learning. Ph.D. thesis, The University of Waikato.

[31] T. Augustine, P. Vasudeva Reddy, P.V.G.D Prasad Reddy, (2018) ‘Intelligent Feature Extraction through Bigram & Trigram Schemes for Detecting Network Intrusions’ International Journal of Engineering Research in Computer Science and Engineering , ISSN : 2394-2320, Vol. 5,Issue 1, January 2018

[32] T. Augustine, P. Vasudeva Reddy, P.V.G.D Prasad Reddy, (2018) ‘Evaluation of Intrusion Detection using a Novel Composite Metric’ International journal of innovative Research in Science, Engineering and Technology, ISSN : 2347 - 6710, Vol. 7, Issue 1, January 2018