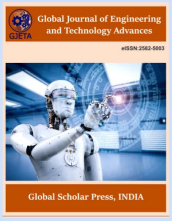
Global Journal of Engineering and Technology Advances
eISSN: 2582-5003
Cross Ref DOI: 10.30574/gjeta
Journal homepage: https://gjeta.com/



(RESEARCH ARTICLE)

Check for updates

Building attack detection system base on machine learning

Rasha Thamer Shawe *, Kawther Thabt Saleh and Farah Neamah Abbas

Department of Computer Science, College of Education, Mustansiriyah University, Baghdad, Iraq

Global Journal of Engineering and Technology Advances, 2021, 06(02), 018–032

Publication history: Received on 09 January 2021; revised on 06 Februay 2021; accepted on 08 Februay 2021

Article DOI: https://doi.org/10.30574/gjeta.2021.6.2.0010

Abstract

These days, security threats detection, generally discussed to as intrusion, has befitted actual significant and serious problem in network, information and data security. Thus, an intrusion detection system (IDS) has befitted actual important element in computer or network security. Avoidance of such intrusions wholly bases on detection ability of Intrusion Detection System (IDS) which productions necessary job in network security such it identifies different kinds of attacks in network. Moreover, the data mining has been playing an important job in the different disciplines of technologies and sciences. For computer security, data mining are presented for serving intrusion detection System (IDS) to detect intruders accurately. One of the vital techniques of data mining is characteristic, so we suggest Intrusion Detection System utilizing data mining approach: SVM (Support Vector Machine). In suggest system, the classification will be through by employing SVM and realization concerning the suggested system efficiency will be accomplish by executing a number of experiments employing KDD Cup'99 dataset. SVM (Support Vector Machine) is one of the best distinguished classification techniques in the data mining region. KDD Cup'99 data set is utilized to execute several investigates in our suggested system. The experimental results illustration that we can decrease wide time is taken to construct SVM model by accomplishment suitable data set pre-processing. False Positive Rate (FPR) is decrease and Attack detection rate of SVM is increased .applied with classification algorithm gives the accuracy highest result. Implementation Environment Intrusion detection system is implemented using Mat lab 2015 programming language, and the examinations have been implemented in the environment of Windows-7 operating system mat lab R2015a, the processor: Core i7- Duo CPU 2670, 2.5 GHz, and (8GB) RAM.

Keywords: Attack Detection0, Intrusion Detection System (IDS)0 ,Data Mining0, Support Vector Machine (SVM)0.

1. Introduction

With the rise utilize the networkeds computerss fors crucial systemss ands thes common utilize of distributed ands large computers networks, the security of computers networks concern rises and network intrusions have been a dangerous risk in latest time. Intrusions detections systems (IDS) hass beens great used to be a seconds row of protection form networked computers systems sideways with additional network security methods for instance access controland firewall. The majoraimof IDS is to detect illegal utilize,abuse and misuse of computers systemss by boths systems insiderss ands outsiders intruders. Theres are differentsmethodssto construct intrusions detections systemss (IDS). IDSs can bes classifieds into twos classifications depend on the approachess utilized to detects intrusions: abuse detectionsand anomalysdetection[1, 2, 3]. Anomalysdetectionsmethodscreatessthe profilessofsusual actions ofs users, system resources, operatings systems, networks services ands traffic usingthe examination trails created by anetwork scanning program or a host operating system. This method detects intrusions by classifying important perversions from the usual attitude samples of these profiles. Anomaly detection method is not necessary that previous knowing of the security holes of the goal systems. So, this system is capable tosdetect not onlysidentified intrusionssbut alsosunidentifiedsintrusions. Moreover,thissmethodscan identify the intrusionssthatsare accomplished by the misuse

* Corresponding author: Rasha Thamer; Email: rashathamer82@uomustansiriyah.edu.iq
College of Education in Computer Science AL-Mustansiriya University Baghdad, Iraq.

of legal users or disguise without violation security politics [4,5,26]. The drawbackssof this methodwere it hadsrise fakespositive recognition fault, the hardness oftre atment progressive misbehavior, and costly calculation [4,6, 7]. otherwise, misuse recognition method determines doubtful abuse signatures depended on known system weaknesses and a security procedure. Abuse method achievesswhether signaturessof identified attackssare existing or notsin thesauditing paths, andsany corresponded behavior is recognized ansattack. Misusesdetection method identifiessonlyspreviously recognized interferencessignatures. Thesbenefit ofsthis method issthatsit seldom defeat to identify prior toldsintrusions,si.e.decrease fake positivesrate [5,s8]. Thesdifficulties of this method cannot identify modern intrusions it have not ever before been detected, i.e. Greater fake negativesrate. Morever, thissmethod hassanother disadvantages as the hardness of misusessignature bases and the hardness of creating and updatingssignature rulessof intrusion [4, 8, 9]. These are twostypessof Intrusionsdetectionssystems IDS are NetworksIntrusion DetectionsSystem NIDSsand Host-basedsIntrusionsDetection (HIDS) [9,10,27]. In our study we built Intrusion detection systems (IDS) based on data mining.

Intrusion detection systems (IDS) approachess are two complementary orientations in intrusion detection [11,12,28]: Misusesdetection. Thesseeking forsproof ofsattacks depend onsthe information collectedsfrom recognizedsattacks.moreover, it is indicate to attacks type as detectionsby appearance or misusesdetection. Anomalysdetection is seeking for perversions from the pattern of uncommon attitude depend on the monitoring of a systemsduring a ordinary status andsis indicated to such anomalysdetection or find via behavior.

2. Related work

2.1. Soni andSharma in 2014[13]

suggested two techniques artificial neural network (ANN) and C5.0 are employed together with characteristic picking. Feature picking method seliminate several irrelevant characteristics while C5.0sand ANNsperformed such a classifiersto categorizesthe input datasin eithersnormal category orattack that onesof the fivestypes. KDD99sdataset is employed to test and train the system; C5.0ssystem through numberssof characteristics is make improvedsresults withsnearly 100% accuracy. Morever, they used ANN approach to categorize intrusion data depend on their partition size. A comparative result demonstrates that C5.0 is execution better than ANN and yields best outcome with 36 features.

2.2. Zargar and Baghaie in 2012[14]

offered a category-basedspicking of active parameterssfor detection of intrusion utilizing PrincipalsComponents Analysiss (PCA). They employ 32smain characteristics from **Transmission Control Protocol// Internet Protocol** (TCP/IP) sheader, also 116sresulted featuressfrom TCPsdump are picked in adataset of networkstraffic. Attackssare classified in fourssets, User attack (U2R), Denialsof Services (DoS), Remotesto and Probingsattack, Remotesto Usersattack (R2L). Moreover, they used TCP dumpsfromsDARPA 1998sdataset insthe tests as the pickedsdataset. PCAsapproach is utilized to define an ideal characteristic setsto produce thesdetection procedure higher speed. The experimentalsresults display that characteristic reductionscan get better detectionsrate for the category-basedsdetection method while the continuing detectionsaccuracy within asuitable range. The KNNsclassification technique is utilized for the attackssclassification. The experimentalsresults illustrate that characteristic reductionswill importantly speedsup the testing and training the time for recognition of thesintrusion challenges.

2.3. Mukkamala and Sung in 2003 [15]

proposed Feature picking for Intrusion Detection utilizing Two learning machine classes for intrusion detection system (IDS) aresstudied: ArtificialsNeural Networkss (ANNs) andsSupport VectorsMachines (SVM). They display that SVMs are better than ANN insthree serious respectssof intrusion detection system: SVM execute and train are greatness quicker; SVMs scale much superior; and SVM provide greater classification accuracy. Moreover, address the concerning matter of rankingsthe significancesof input characteristics, whichsis a hazered of major significant. Sincesremoval of the useless and/orsunimportant inputssproduce a simplified hazered and probably quicker and extra precise detection, characteristic picking is quite significant in intrusion detection. The experimental results show that SVM-depend IDS utilizing a reduced feature numbers can deliver improved or comparable performance. In conclusion, IDS suggesteddepend on SVM for detecting an exact category.

2.4. Zhu et al.,2005 [16]

RICGAs (ReliefFsImmunesClonal GeneticsAlgorithm), a collective characteristic subsetspickingmethod depend on the Immune Clonal selection,ReliefFsalgorithm and GA is suggested in isemployed BP networks as classifier. RICGA has

higher accuracy of classification (86.47%)forssmall scope characteristic subsetssthansReliefF-GA. In the paper, the features are not mentioned.a composite characteristic subset selection method, claimed RICGAs (ReliefFsimmune clonalsgenetic algorithm), depend on the immune clonal selection algorithm ,ReliefF algorithmand GA. In the RICGA method, they employed in the first ReliefFsto getsrid ofsirrelevant characteristics, then execute a improved geneticsalgorithm to get the lastly characteristicssubset. They analyse hardly theRICGA Markovschain modelsalgorithm and itssconvergence. Experimentalsresults on realsKDD CUP'99 datasets display that the RICGAsmethodsis eminent to the ReliefF-GAsand GAson classificationsprecision and input characteristic subsetssize.

2.5. Ming-Yang Su (2011) [17]

offered anapproach for featurespicking to identify DoS/DDoSsattacksfor designing in realstime ansanomaly-based Network Intrusion Detection System NIDS. Genetic algorithms (GA) collective with KNN (knearest-neighbor) are utilized for characteristic weighting and picking. The outcome of KNNsclassification is employedas the fitnesssfunction in a GA to improve the characteristics weightsvectors. First 35scharacteristics in the train stage aresweighted. The highest 19 characteristics are taking into account for recognized attackssand the tops28 characteristics for unidentifiedsattacks. In this paper, extractedscharacteristics are not aforesaid. Atotal accuracy rateof 97.42% is obtained for recognized attacks and 78% for unidentified attacks.

3. Data Mining and Intrusion Detection Systems

Intrusion0detection systems0(IDS) have been depend intraditionally on0the feature of an attack and the system tracking activity to check if it matches that description. IDS depend on data mining is creation their0appearance more ability. The system of Data0Mining approaches for0intrusion detection applications have been commonly employed these days. The0intrusion detection difficult has been0reduced to a0Data Mining mission of classifying data.Summarized, given a data pointsset belonging to various attacks activity (0Classes0) and purposes to isolate them assaccurately as possible by means of0a model0. Many various data0mining approaches found for intrusion detection0classification. In this system, we employed a Support Vector Machine (SVM) for attach detection as a classification algorithm. Also, we used feature extraction and dimensionalitysreduction algorithmss(PCAsand LDA, SVDs) basing on the KDD'99 Cupsdatasets.

4. Design and Implementation of Proposed System

The proposed intrusion detection system to scheme a proficient intrusion detection and recognition system is described as follows:

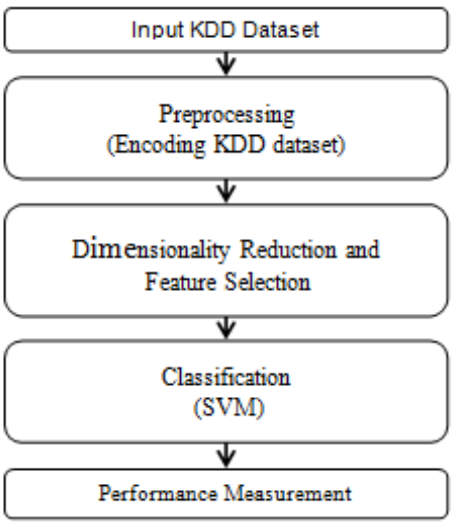


Figure 10 Intrusion Detection Classification proposed System**0**approach

The aim of analyses is to increase the intrusion detection system achievement; the data which used as input to proposed systemis KDD Cup 99 dataset. The KDD Cup 99 datasetis requirement to pre-processing which is done by converting all data into similar format. Then feature reduction is performed to extract and reduction features.Finally, intrusion classification stage is done by based on different kind of system insertions, the classification algorithms Support Vector

Machine (SVM). As KDD Cup 99 dataset holds some symbolic attribute and also numeric attributes, two sorts of transformation technique have been utilized for these properties. The two machine learning procedures are prepared on both kind of transformed dataset and afterward their outcomes are looked at with respect to the correctness of intrusion detection. The suggested system is containing fundamentally of two essentialjobs which aresfeature reductionsand attack detection.

Our proposedsintrusionsdetection system steps are showed in Figure(1), swhich includes the main parts. Input KDD'99 Dataset, Dataset Pre-processing, Dimensionality Reduction and feature selection, Classification0 Algorithms0 and Performance0Measurement.

4.1. KDD'99Input Dataset

In first phase of the suggested intrusion detection system gets the KDD Cup 99dataset as an input where the whole record numbers. In our proposed system, we utilized the total KDD Cup 99 dataset. Each record on 42 features; the records have labeled either attackor normal type.

The KDDsCUP 1999 [18] standard datasetssare employed tosevaluate various characteristic selectionstechnique for Intrusionsdetectionssystem. This system contains of 4,940,000srelatedsrecords. Every relation had a labelsof eithersattack or the normal kind, with quite one exact attack category happens in one of the four attacks types [19] as: User to Root Attacks(U2R), Remote to Local Attacks(R2L),Denial of Service Attack (DoS) and Probing Attack.

Denial of Service Attack (DOS): Attackssof this category deprivesthe legitimate or host user from utilizing the resources orservice.

Probe Attack: Thesesattacks mechanically scan a computer networks or a DNS server to getlegal IPsaddresses.

Remote to Local (R2L) Attack: In thissattack category an attackerswho doessnot have an accountson a victim machine achievements localsaccess to the mechanism and changes the data.

User to Root (U2R) Attack: In this attackscategory a localsuser on a mechanism is ablesto get excellence normally kept for the supers (root) users. Each related recordscontained of 41 featuressandsalso are labeled in configuration ass1,2,3,4,5,6,7,8,9,.....,41 and fallssinto the fourstypes are displayed inTable 1:

Category 1 (1-9): Elementary featuressof separate TCP associates.

Category 2 (10-22): implicate featureswithin an association proposition by domainsknowledge.

Category 3 (23-31): Trafficsfeatures calculated utilizing a two-secondstimeswindow.

Category 4 (32-41): Trafficsfeatures calculatedutilizing a two-secondstime windowsfrom goal to host.

Table0 **1 . Distribution**0 **of**0 **intrusion**0 **Types**0 **in**0 **datasets**0

Dataset0	Normal00	Probe00	DOS00	U2R00	R2l00	Total00
KDD cup data of 10% S	97280	4107	391458	52	1124	494020

4.2. KDD'99 Pre-processing

KDD'99 pre-processing is second phase and is one of the significant phases of system. This stage proper data to be accepted to next phase for extraction and reduction data. This phase contain from two step(Dataset Labeling, Normalization). The following subsection will illustrate all details about these steps:

4.2.1. Dataset Labeling

The Dataset Labeling is the first step in KDD'99 Pre-processing phase. This step is so important. The output of Dataset Labeling employed as input to next step in Pre-processing phase (Normalization). The dataset labeling is done by utilizing the whole features in thes KDD 10%0corrected0datasets at it displayed in the screen shot which is sited in the

second cell of the entire dataset. The figure (2) isthe KDD 99 dataset screenshot that we took it from environment of our matlab.

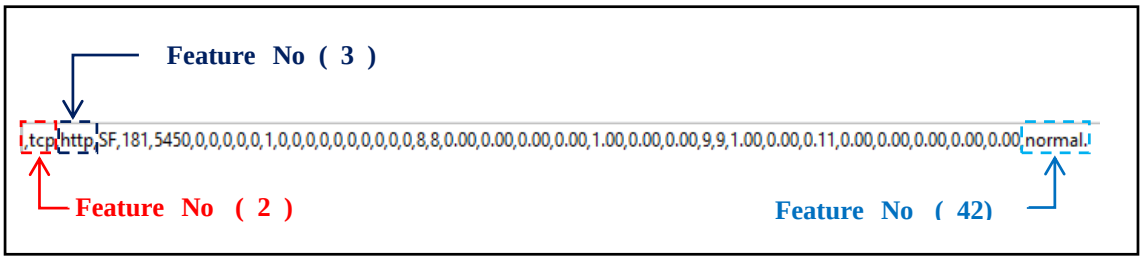


Figure 2 First KDD cup dataset row of 10% correction (data sample)

So, the Table (2) illustrationlabeled the dataset base onattacks which0are fall into one of fives categoriessas belows:

Table02 Our Class labeling of "10% KD099" dataset

Attackss Types	Descriptions	Subs Typess	Labels
(0DoSs0) Denials of sService	Attackers triess to prevents legitimates userss froms using as services	Smurfs	1
		Neptunes	
		backs	
		Teardrops	
		pods	
		lands	
Normals	datas withs nos attacks	normals	2
Probess	Attackers triess to prevents legitimates userss froms using as service.	Satans	3
		ipsweeps	
		portsweeps	
		nmaps	
(R2L) Remotes tos Locals	Attackers does not have an accounts on the victim machines, hence tries to gains accesss	warezclients	4
		guess_passwds	
		Warezmasters	
		Imaps	
		ftp_writes	
		Multihops	
		Phfss	
		spys.	
Users to Roots (U2R)	Attackers has local accesss to thes victims machines and tries to gains supers useres privilegess	buffer_overflows	5
		Rootkits	
		Loadmodules	
		Perls	

The datasets records0includes 42 characteristics (e.g0,, 0service0, protocolstype0, andsFlag) and is labeledsasseither attack or a normal also illustrateany one of attack type as presented in Figure (1.2).as an example,if we take a sample of first rowfromthe KDD 99 dataset before doing the normalization. The Figure (2) is clearthat the feature numbers is (42) which has the definite attack categoryas describedin Table (2).

The Table (3) illustrate the dataset how must be labeledsby employing 10% ofsthe correctedsdataset:

Table 3 nominal features after Label

Types	**Features Names**	**Numeric0 values**
Protocol-type**s**	TCP0	10
	UDP	20
	ICMP0	30
Flag**s**	SF0	10
	S10	20
	REJ0	30
	S20	40
	S00	50
	S30	60
	RSTO0	70
	RSTR0	80
	RSTOS00	90
	OTH0	100
	SH0	110
Service**s**	All services00	10 to 660
Attack**s**	All attack00	1 0to 230

There is also another issue in this step. There are many nominal values in the dataset such as HTTP, SF, and ICMP. Consequentlyin this step transformall nominal values to numeric values in advance. For instance, the service form of "tcp" is mapped to 1,"udp" is mapped to 2,"icmp" is mapped to 3 and the table (3) shows all transformation the dataset nominal value features into the numeric values. Figure (3) has shown the original KDDCUP1999 dataset will become after transformation as display in figure (3).

Before labeling

0,icmp,ecr_i,SF,1032,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,511,511,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,255,1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,smurf.

0,tcp,http,SF,248,2129,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,23,23,0.00,0.00,0.00,0.00,1.00,0.00,0.00,23,255,1.00,0.00,0.04,0.03,0.00,0.00,0.00,0.00,normal.

After labeling

0,3,8,1,1032,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,511,511,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,255,1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,6.

0,1,1,1,248,2129,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,23,23,0.00,0.00,0.00,0.00,1.00,0.00,0.00,23,255,1.00,0.00,0.04,0.03,0.00,0.00,0.00,0.00,1.

Figure 3 Pre-processing Original KDDCUP1999 dataset before and after transformation

4.2.2. Normalization

After we do the labeledfor all dataset feature space, we can do the dataset Normalization by using the whole KDD010% corrected0 datasets at it shown0in the screens shots which is located0in the second cell of the wholes datasets.

KDD'99 as an input dataset includes characteristic numbers and theses are in0different style. Somesare numbers of style and others are in character style. Consequently, in this stage this various style dataset is transformed into samestyle to be extracted0to thes next phases.

Sinces theres are some KDD CUP 99 dataset features are continuous, therefore a normalize process is done on these features to become more suitablefor the DM classification algorithms. Normalization is utilized for preprocessing the data, where the characteristic data arerangeas to be in a tiny definite scaled for example 0.00 to 1.00 or-1.00 to 1.00. Normalizing0the0input values for every characteristic measuredsin thestraining patterns willsaid speedsup the learningsphase.

4.3. Features Extraction0 and Dimensionality0 Reduction 0of the0 KDD990

Features extraction0and dimensionality0reduction method is done by eliminatin gredundant and irrelevant features. Irrelevantis that features have little connection with class labels. The redundant features have robust relationship with picked features. in this suggested system we employed three various algorithms which are Singular value decomposition (SVD), Principal Component Analysis (PCA),and Linear Discriminant Analysis (LDA) approaches. We used these techniques for extracting appropriate characteristics from dataset, and reduce dimensions of the KDD as wellthen are given as an input to a next step.

4.3.1. Principal Component Analysis (PCA)

PCA is a convenient statistical approach that has found systems in fields for instance image compression and face recognition, and is a popular method for definition samples in high dimension data. Theswhole object ofsstatistics is depend on about the conception which big data set, and examine set describes of the relations betweensthe separate pointssin thatsset [20].The objective of PCAsis to limit the dimensionalitysof the dataswhile preserving as far as probable of the variance current in thesoriginalsdataset. It is asway of categorizing samples insdata, and term the data in as a technique as to focus their differences and similarities [21].

Algorithm1 Principal Component Analysis (PCA)

Algorithm:

Suppose $x_1, x_2, \ldots x_M$ are $Nx1$ vectors

Step 1: $\overline{x} = \frac{1}{M}\sum_{i=1}^{M} x_i$

Step 2: Subtract the mean: $\Phi_i = x_i - \overline{x}$

Step 3: From the matrix $A = [\Phi_1 \Phi_2 \ldots \Phi_M]$ (N *M matrix), then compute:

$$C = \frac{1}{M}\sum_{N=1}^{M} \Phi_n \Phi_n = AA^T$$

(Sample covariance matrix, N*N, characterizes the scatter of the data)

Step 4: Compute the eigenvalues of C: $\lambda_1 > \lambda_2 > \cdots > \lambda_N$

Step 5: Compute the eigenvectors of C: $u_1, u_2, \ldots u_N$

Since C is symmetric, $u_1, u_2, \ldots u_N$ form a basis, (i.e., any vector x or actually $(x - \overline{x})$, can be written as a combination of the eigenvectors):

$$(x - \overline{x}) = b_1 u_1 + b_2 u_2 + \cdots + b_N u_N = \sum_{i=1}^{N} b_i u_i$$

Step 6: (dimensionality reduction step) keep only the terms corresponding to the K largest eigen values:

$$x - \overline{x} = \sum_{i=1}^{K} b_i u_i$$

How to choose the principal components?

- To choose K, use the following criterion

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} > \text{Threshold (e.g., 0.9 or 0.95)}$$

4.3.2. Singular-Value Decomposition(SVD)

Another approach we use it in this system which is Singular-Value Decomposition (SVD). the figure(4) explain the form of a SVD,Let X be an m × n matrix, and let the rank of X be r. The rank of a matrix is the biggest number of rows "or equally columns" we can selectfor which numberof non zero linear set of the rows is the vector 0 (all-zero) in this case a set of such columns or rows is independent. Also, the Figure (1.4) displays the matrices U, Σ, and V as with the following properties:

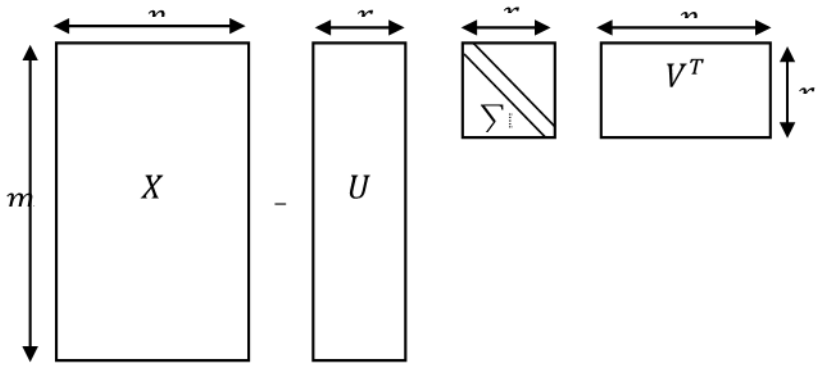


Figure 4 The form of a singular-value decomposition

1. U is an $x \times r$ column0S-Sorthonormal matrix0S; that0is, each of its0Scolumns is a unit0 vector and the dot product of any two column0S is 0.
2. V is an $n \times r$ column0-orthonormal0 matrix0, note0 that we0 always use V in its transposed form0, so it is the rows of V^T that are0 orthonormal0.
3. Σ is a diagonal0 matrix; that0 is, all elements0 not on the main diagonal0 are 0. The elements of Σ are called0 the singular0 values0 of X.

Singular0-ValueS Decomposition0S (SVD) algorithm0S steps are described0S in the algorithm0 (2) below0

Algorithm0 **(**02**0)** 0 **Singular**0**-Value**0 0**Decomposition**0 **(**0**SVD**0**)**

Input0**:** Generate0 Data0 matrix0 $\boldsymbol{X}$
Output0**:** New0 Dimensions $\boldsymbol{C}$

1. **Repeat**0
2. Applying0 SVD to the matrix0 X as $X = USV^T$
   $X \rightarrow$ is an $m \times n$ matrix
   -$m \rightarrow$ no. of sessions0 (0vectors0)

-$n \rightarrow$ is no. of TH attributes)
$U \leftarrow XX^T$matrix0 of 0the eigenvectors0
S is matrix0S which0 is diagonal0
$V \leftarrow$ is matrix the0 eigenvectors0.

3. Construct0 the0 covariance matrix0 from 0this 0decomposition by0
   $$XX^T\ XX^T \leftarrow (USV^T)(USV^T)^T = (USV^T)(VSU^T)$$
4. $V \rightarrow$ an 0orthogonal matrix0 $(V^TV = I), XX^T = US2U^T$
5. square 0roots of the eigenvalues 0of $X\boldsymbol{X^T}$ are the singular0 values0 of X
6. **until**Represent0 every 0transaction Ii over0 the time interval0 t as a0 vector0 $0x(t)_i$
1. **Return**0 U^TX

4.3.3. Linear discriminant analysis

Linear discernment analysis (LDA) is a different technique that employed for reduction of dimensionality and feature extraction. LDA requires reducing dimensionality while maintain as much of the class distinctive information.

LDA algorithm phases are presented in. LDAsis a high-dimensionalsdata analysis approach and appropriate for characteristics transformationsto ease classifications [22]. Thereshas beensa capability to utilize PCAsmethod for characteristics subset piking or reduction in many various applications such facesrecognition, text recognition and

handswritten, and image compression, in addition to intrusionsdetection [23] but LDAshas more advantages over PCAsand is favored over PCAsowing to thesfollowing aims.

a) LDAsoutperforms PCAsin example of great numbersof samplesdataset [24].

b) LDAsdirectly treats withsboth differentiation within-classessas well assbetween-classes whilesPCAsdoes not havesanysconception ofsthe between-classessstructure [25].

c) LDA maintenance class differentiation informationsas muchsas probable while accomplishment dimensionalitysreduction [9].

Algorithm 3 A1gorithm for Linear Discriminant Analysis

LDA0 Algorithm0 Steps0

Suppose**0** X=($X_1X_2X_3X_4$……X_C) are Nx1 features vectors0s where C and each features vectors contains ns feature samples. Followings are steps adapted ins LDA algorithms.

Steps0 10: Computes thes between class scatters using completes features samples.

$$S_b = \sum_{i=1}^{C} (\alpha_i^j - \alpha_i)(\alpha_i^j - \alpha_i)^T$$

Step 2 Calculates the Total classs scatter matrixs

$$S_t = \sum_{i=1}^{C} \sum_{j-1}^{n} (\alpha_i^j - \bar{\alpha})(\alpha_i^j - \bar{\alpha})^T$$

Step 3 Computes Eigenvaluess and Eigenvectors using Eigens equation fors LDA.

$S_T X = \lambda S_i X$

Step 4 Computes the Eigenvectorss corresponding to Eigenvalues such that and Eigenvectors: X1, X2, X3 … XN where N represents dimensionality of feature vector and N in our case

$$Eigenvalues: \lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots\ldots\ldots \lambda_N$$

Step 5 Evaluate the contribution of each feature vector

$$C_j = \sum_{p=1}^{m} |V_{pj}|$$

Step 6 Sort the features vectors in descending orders corresponding to their impact or contributions.

Step 7 The dimensionality reductions phase based on largest0 eigenvalues is skipped as the selection of optimum subsets of linear components

4.4. ClassificationSupport Vector Machine Algorithms (SVM)

Support Vector Machine (SVM) is a machine learning method setsem ployed for regression and classification. SVM is depending on the idea of decisionplanes that describe decision boundaries. A decision plane is one that splits between a set of matters having various class memberships.Our suggested insertion detection system depend on dimensionality reduction which PCA and SVD, LDA algorithm which has employed one classification outcomes.

4.5. Performance Evaluation

The insertion detection system efficiency is evaluatedsby its capacity tosmake precise estimates. Accordingsto the realsnature of a grant eventscompared tosthe forecast fromsthe IDS, four probable results are presented insTable(4),

famous assthe confusionsmatrix [4]. Detection Rates (DR) or True Negative Rates (TNR),True Positive Rates (TPR), FalsesPositive Rates (FPR) or FalsesAlarm Rates (FAR) and FalsesNegative Rates (FNR) are gauges that cansbe practical to quantifys the execution of IDS [4] depend ons the above confusionsmatrix

Table 4 Confusion Matrix

Predicted0 Actuals	0Negative0class (Normal)	0Positive0Class (Attack)
0Negative0class (Normal)	True0 Negative0 (TN)	0False Positive0ss (FP)
0Positive0Class (Attack)	False0 Negatives (FN)	Trues Positive0 (TP)

We has gotten accuracy by recognition0rate is illustrate such as the ratio0between thes correct recognition numbers decisionsto the number0of total.

$$Accuracy = \frac{TP + TN}{Total\ number\ of\ test\ samples} * 100$$

5. Results and discussion

Tables (5) and (6) displays the overall performance results of Support Vector Machines(SVM) on KDDsCup 99 dataset based on testingsand training by utilizing three various algorithms (0PCA and LDA, SVD0) that0 we have0 offered in our system0.

Table 5 Accuracyusing (SVM) and different feature extraction and dimensionality reduction algorithms on the training dataset

Dataset0Features	**Classification Algorithm**	**Dimensionality0Reduction Algorithm0s**		**Accuracy0**
		Algorithm0	Feature 0No0.	
0420	SVM	PCA	07	98.43943
042	SVM	LDA	04	99.24772
042	SVM	SVD	07	98.7197

Table0 6 Accuracy**0** fors using (SVM) with**0** different dimensionality**0**reductions algorithms on the testing dataset

Datasets Features 0	**Classification Algorithm**	**Dimensionality0Reduction0 Algorithms**		**Accuracy0**
		Algorithms	Features No0.	
042	SVM	PCA	07	91.4771
042	SVM	LDA	04	99.4511
042	SVM	SVD	07	98.0556

Figure (5) explains the performance results on the training dataset employing three dimensionality reduction algorithms. For attack detection we utilized support Vector Machine classification algorithm.

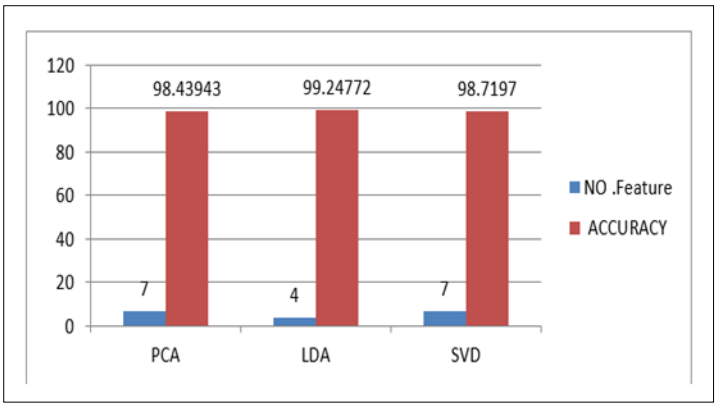


Figure 5 Accuracy of Support vector machine (SVM) classification with three Algorithms Dimensionality Reduction Attack Detection

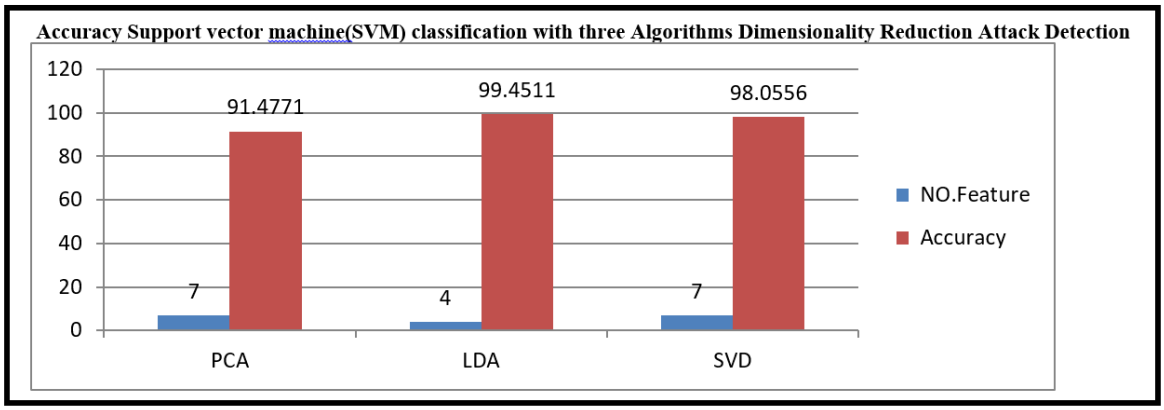


Figure 6 Accuracy Support vector machine (SVM) classification with three Algorithms Dimensionality Reduction Attack Detection

5.1. Experiential Results using the Whole Dataset Samples

Big data analysis is a major change theses day, so in term of dealing with a huge number of data samples (records).In our proposed system we design anapproach to transact with intrusion detection classification difficult. Also we used a huge number of data examples (494,201) with entire feature numbers (42 features). To execute and solve the problem of 10% KDD classification, we propose a data folds segmentation. In this part, we tried to display the experimental results employing the entire data samples which are (494,201). These experimental results of intrusion detection proposed system which is the intrusion0detection0classifications systems depend on various features reductions algorithms0 on the KDD Cup 99 dataset.

Each record of dataset labels includes one of the 5 type of attacks. Since the 494,201 is a big data analysis especially in our proposal which we employed three various algorithms to do the dimensionality reduction and feature selection. In each one we used for0reducing the042 features of the(0 KDD data sets) and two classification0 algorithmss to detect the four types of IDs attackss.

Our methodology proposed system for dealing with this type of big data analysis is to divide the entire dataset sample (494,201) to k-folds (k-folder). Each fold (folder) has n-sample numbers from the dataset. Each sample has been picked in term that no fold (folder) has the same data sample as another fold.

In experimental result of proposed system, we divide the dataset to 25-folds. The 24th folds, each one has 20000 data samples, and the last one has 14021 samples so all these shown in Table (7) below.

Table 7 Whole data sample approach

Total Dataset size	**Data approach 25-folds splitting**	
	1 to 24th fold	25th fold
494,201	20000	14021

The experimental results were examined and discussed to exemplify the proposed ID system. In this case, we described three major parts. The first part is the essential features by employing three algorithms that have selected from the entire feature space which is 42 features. The second part, we explained in this step the result of dimensionality reduction imprudent and feature selection algorithm by selecting the feature space (7). The last part is a part of comparing between IDS proposal experimental results and the previous works.

In this implementation system,we trust on scoring the Eigen value score to reorder the feature from the highest score (the most significant one) to the smallest one

5.2. Classification Experimental Results

classify the kinds of attackon the 10% of KDD Cup 99 dataset, we employed. Support Vector Machine (SVM) classification algorithm in this phase. This algorithm has been utilized with the reduction dimension space features.

5.3. Comparing our Classification Results

Compare the performance results of SVM classifiers by employing the whole data samples with all three dimensionality reduction methods that we have suggested.

Table (8) displays the performance results of the SVM employing the whole data sample basing on the utilizing the all dimensionality reduction methods.

Table 8 A Compression results for Insertion Detection using SVM

Dimensionality Reduction Algorithm	**Dimensionality**	**Training**	**Testing**
PCA	21	94.08%	93.79%
	11	94.10%	93.82%
	7	94.13%	93.83%
LDA	4	92.28%	98.11%
SVD	21	93.36%	91.65%
	11	93.44%	91.75%
	7	91.77%	90.12%

We can see that by employing the SVM classifier to categorize the entire data samples according to the attack kinds, our method for PCA and LDA, SVD gives a higher accuracy in testing and training.

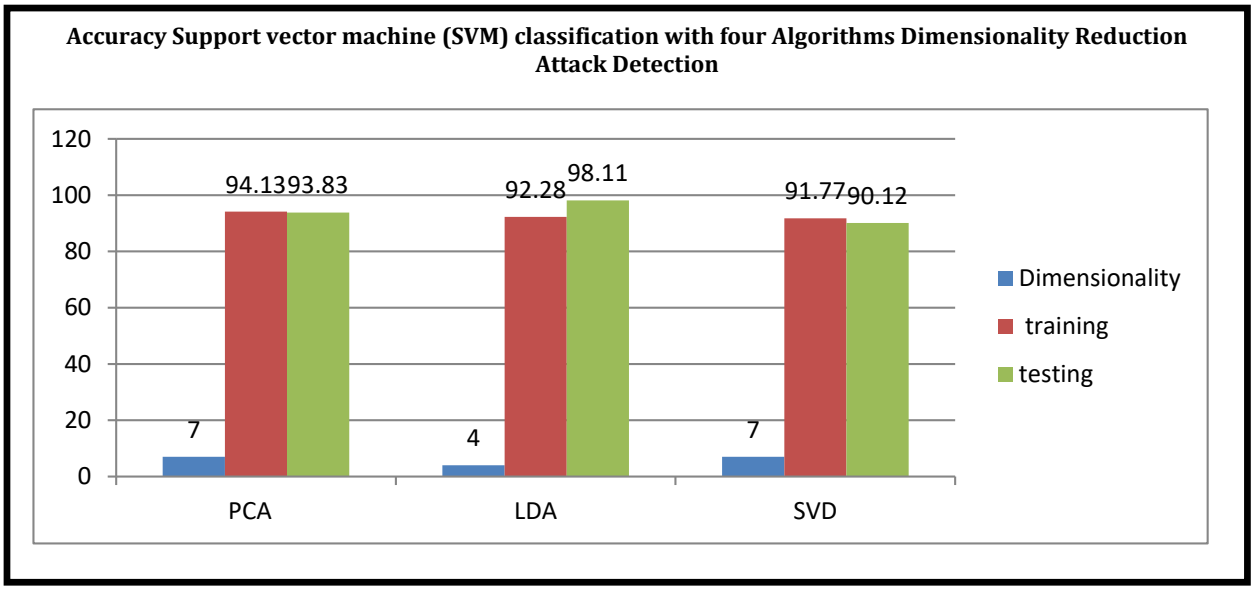


Figure 7 Accuracy Support vector machine (SVM) classification with three Algorithms Dimensionality Reduction Attack Detection

5.4. Comparing our Classification Results with Other studies

There were great number studies that have been done to classify the attack kinds using 10% of KDD Cup 99 dataset. In this part, we will compare our results utilizing reduction algorithm (PCA and LDA, SVD) with the other studies that have been done on the same dataset. Table (9) shows a briefly comparison between our proposed system's result and the other methods according to the performance results for the overall accuracy for testing and training.

Table 9 A Compression results for Insertion Detection between the previous studies and our approach

Approach	Classification Alg.	Feature No.	Accuracy
Soni and Sharma (2014)	C5.0	32	99.49%
	ANN	32	99.49%
Zargar and Baghaie (2012)	KNN	42	99.41%
	KNN	Effective features are used from (42)	96.70%
Mukkamala and Sung (2003)	ANN	41	87.07%
	ANN	34	81.57%
(Zhu et al.,2005)	BP Network	12	88.15%.
Ming-Yang Su (2011)	GA KNN	19 28	97.42% 78.00%
Suggestion	SVM	7	98.00%
	SVM	4	99.45%

6. Conclusion

Today, a large amount of threat attacks network and information security. In this paper, we proposed an intrusion detection system that reduces the set of features and classifies attack types. The reduction of features is performed by us also then the classification which the proposed algorithm is a combination of features selection. Reduced features for intrusive detection system and increased attack detection rate to the SVM applied classification algorithm, which gives the highest resolution. The cup1999 kdd selection attacks are identified with less Error rate and high accuracy. The feature selection and their reduction have both affected the performance of the classification algorithm. In the future swarm optimization function dynamically reduces the number of unused feature attribute of traffic data.

Compliance with ethical standards

Acknowledgments

The authors would like to thank the Mustansiriyah University (www.uomustansiriyah.edu.iq) Baghdad, Iraq, for supporting this work.

Disclosure of conflict of interest

All authors declare that they have no conflict of interest0.

References

[1] Axelsson, S, "Intrusion Detection Systems:A Taxonomy and Survey," Technical Report No 99-15, Dept. of Computer Engineering, Chalmers University of Technology, Sweden, March 2000.

[2] Lunt, T F, "Detecting Intruders in Computer Systems," in proceeding of 1993 Conference on Auditing and Computer Technology, 1993.

[3] Sundaram, A, "An Introduction to Intrusion Detection," The ACM Student Magazine, April 1996; Vol.2, No.4,. Available at http://www.acm.org/crossroads/xrds2-4/xrds2-4.html

[4] Porras, P A, "STAT: A State Transition Analysis Tool for Intrusion Detection," MSc Thesis, Department of Computer Science, University of California Santa Babara, 1992

[5] Dorothy E. Denning, "An Intrusion Detection Model," In IEEE Transactions on Software Engineering, February 1987; Vol.SE-13, Number 2: page 222-232.

[6] Lunt, T F, et al., "A Real-time Intrusion Detection Expert System (IDES)," Technical Report SRI-CSL-92-05, Computer Science Laboratory, SRI International, Menlo Park, CA, April 1992.

[7] Mykerjee, B, Heberlein, L T and Levitt, K N, "Network Intrusion Detection," IEEE Network, 1994; Vol.8, No.3:pp.26-41.

[8] Ilgun, K, Kemmerer, R A, and Porras, P A, "State Transition Analysis: Rule-Based Intrusion Detection Approach," IEEE Transactions on Software Engineering, March 1995; Vol. 21, No. 3:pp.181-199.

[9] Kumar, S, "Classification and Detection of Computer Intrusions," PhD Thesis, Department of Computer Science, Purdue University, August 1995

[10] A Macgregor, M Hall, P Lorier and J Bruskill, "Flow clustering using machine learning techniques", In PAM 2004, Antibes-Juan-Les-Pins, France, LNCS. 2004; pp. 205-214.

[11] S. Kumar, Classification and Detection of ComputerIntrusions, Ph.D. Thesis, Purdue University.

[12] MithcellRowton, Introduction to network security intrusion detection, December 2005.

[13] p Soni , p Sharma ," An Intrusion Detection System Based on KDD-99 Data using Data Mining Techniques and Feature Selection" ,International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, July 2014; Volume-4 Issue-3,

[14] G R Zargar and T Baghaie,"Category-Based Intrusion Detection Using PCA",Journal of Information Security, 2012; 3:259-271 http://dx.doi.org/10.4236/jis.2012.34033 Published Online October 2012 .

[15] S Mukkamala, A H Sung ," Feature Selection for Intrusion Detection using Neural Networks and Support Vector Machines" ,Transportation Research Record Journal of the Transportation Research Board · January 2003 DOI: 10.3141/1822-05.

[16] Zhu, Y. et al. Modified Genetic Algorithm based Feature Subset Selection in Intrusion Detection System. Proceedings of ISCIT 2005; 9-12

[17] Ming-Yang Su Real-time anomaly detection systems for Denial-of-Service attacks by weighted k-nearest-neighbor classifiers. Expert Systems with Applications, 2011; 38:3492–3498

[18] sKDD Cup 1999 Intrusion detection dataset: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[19] Mukkamala, S. et al. (2005). Intrusion detection using an ensemble of intelligent paradigms.Journal of Network and Computer Applications, 28(2), 167–82.

[20] Lindsay I Smith A tutorial on Principal Components Analysis February 26,2002.

[21] S. Gong et al., Dynamic Vision: From Images to Face Recognition, Imperial College Press, 2001 (pp. 168-173 and Appendix C:Mathematical Details, hard copy

[22] KresimirDelac, MislavGrgic and Sonja Grgic, "Independent Comparative Study of PCA, ICA, and LDA on the FERET Data Set," University of Zagreb, FER, Unska 3/XII, Zagreb, Croatia. 2006.

[23] M. Turk and A. Pentland, "Eigen faces for recognition," J CognNeurosci 3 (1991), 71–86.

[24] P. Belhumeur, J. Hespanha, and D. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, Proc Fourth EurConf Computer Vision, Vol. 1, 1418 April 1996, Cambridge, UK, pp. 45–58.

[25] A. Martinez and A. Kak, "PCA versus LDA", IEEE Transactions on Pattern Analysis and Machine Intelligence," vol. 23, no. 2, pp. 228-233, 2001.

[26] Putra Wanda , Huang Jin Jie" A Survey of Intrusion Detection System"International Journal of Informatics and Computation (IJICOM)Vol.1, No.1, August, 2019ISSN: 2685-8711

[27] R. Vinayakumar et al." Deep Learning Approach for Intelligent Intrusion Detection System" Received December 27, 2018, accepted January 3, 2019, date of current version April 11, 2019.Digital Object Identifier 10.1109/ACCESS.2019.2895334

[28] H. Alqahtani et al," Cyber Intrusion Detection Using Machine Learning Classification Techniques" © Springer Nature Singapore Pte Ltd. 2020 N. Chaubey et al. (Eds.): COMS2 2020, CCIS 1235, pp. 121–131, 2020. https://doi.org/10.1007/978-981-15-6648-6_1