A novel method for network intrusion detection based on nonlinear SNE and SVM

Yasir Hamid, Ludovic Journaux, John Aldo Leea, M Sugumaran

A novel method for network intrusion detection based on nonlinear SNE and SVM

Yasir Hamid*

Department of Computer Science and Engineering,
Pondicherry Engineering College,
Puducherry, India
Email: bhatyasirhamid@pec.edu
*Corresponding author

Ludovic Journaux

LE2I UMR6306,
CNRS,
Univ. Bourgogne Franche-Comté,
AgroSup Dijon, France
Email: ljourn@gmail.com

John A. Lee

Univ. catholique de Louvain,
UCL/SSS/IREC/MIRO,
Bruxelles, Belgium
Email: john.lee@uclouvain.be

M. Sugumaran

Department of Computer Science and Engineering,
Pondicherry Engineering College,
Puducherry, India
Email: sugu@pec.edu

Abstract: In the case of network intrusion detection data, pre-processing techniques have been extensively used to enhance the accuracy of the model. An ideal intrusion detection system (IDS) is one that has appreciable detection capability overall the group of attacks. An open research problem of this area is the lower detection rate for less frequent attacks, which result from the curse of dimensionality and imbalanced class distribution of the benchmark datasets. This work attempts to minimise the effects of imbalanced class distribution by applying random under-sampling of the majority classes and SMOTE-based oversampling of minority classes. In order to alleviate the issue arising from the curse of dimensionality, this model makes use of stochastic neighbour embedding a nonlinear dimension reduction technique to embed the higher dimensional feature vectors in low dimensional embedding spaces. A nonlinear support vector machine with a radial basis function on a series of gamma values was used to build the model. The results demonstrate

that the proposed model with the dimension reduction has higher detection coverage for all the attack groups of the dataset as well as the normal data. Results are evaluated on two benchmark datasets KDD99 and UNSW-NB15.

Keywords: accuracy; classification; dimension reduction; intrusion detection; KDD99; NLDR; SNE; SVM; UNSW-N15.

Reference to this paper should be made as follows: Hamid, Y., Journaux, L., Lee, J.A. and Sugumaran, M. (2017) 'A novel method for network intrusion detection based on nonlinear SNE and SVM', *Int. J. Artificial Intelligence and Soft Computing*, Vol. 6, No. 4, pp.265–286.

Biographical notes: Yasir Hamid received his Master's degree in Computer Applications from University of Kashmir in the year 2014. He is currently a PhD scholar in Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry. His areas of interests are machine learning, network security, non-linear dimension reduction and data visualisation.

Ludovic Journaux received his PhD in Image Processing and Computer Sciences from the University of Burgundy (France) in 2006. He is currently working as Associate Professor at Agrosup Dijon and is a member of LE2I laboratory (UMR 6306): Laboratory of Electronics, Computer Sciences, and Images. His research interests include image processing, data mining, statistical analysis, artificial intelligence and classification.

John A. Lee received his MSc in Applied Sciences (Computer Engineering) in 1999 and his PhD in Applied Sciences (Machine Learning) in 2003, both from the Université catholique de Louvain, Belgium. His main interests are dimensionality reduction, intrinsic dimensionality estimation, clustering, vector quantisation, and various aspects of image processing. He is a member of the UCL Machine Learning Group and a Research Associate with the Belgian F.R.S.-FNRS (Fonds National de la Recherche Scientifique). Together with Michel Verleysen, he wrote a monograph entitled 'Nonlinear dimensionality reduction' published by Springer in 2007. His current work aims at developing specific image enhancement techniques for positron emission tomography in the center of molecular imaging, radiotherapy, and oncology (MIRO).

M. Sugumaran received his MSc in Mathematics from University of Madras in 1986, MTech in Computer Science and Data Processing from Indian Institute of Technology, Kharagpur, India in 1991, and obtained his PhD from Anna University, Chennai in 2008. He is currently working as Professor and Head of Computer Science and Engineering at Pondicherry Engineering College, India. His areas of interests are theoretical computer science, analysis of algorithms, parallel and distributed computing, and spatial-temporal data.

1 Introduction

The internet in addition, to represent a revolution within the ability to exchange and communicate data has conjointly provided a bigger chance for the disruption and sabotage of knowledge previously thought to be secure (Hernández-Pereira et al., 2009).

One such group of actions that compromise confidentiality, integrity and availability of the system is intrusion and the task of identifying such malevolent demeanour is called intrusion detection (ID), and the device against whom the responsibility of discriminating between legitimate and illegitimate data is called intrusion detection system (IDS) (Stallings, 2007). An IDS passively monitors the network traffic for the possible intrusions. Based on the scope of surveillance and the point of placement in the system an IDS is classified as host-based IDS (HIDS) or network-based IDS (NIDS) and based on the detection methodology an IDS is classified as misuse detection or anomaly detection (Axelsson, 2000). Automatic NID based on machine learning (ML) techniques has been a provenance of great interest in the research community. Directed towards enhancing the detection coverage of IDS a variety of ML techniques have been used extending both supervised and unsupervised techniques. The inherent problem with the prominently used datasets for NIDS, i.e., KDD99 and UNSW-NB15 data is lower detection rate for less frequent attacks. The reason for this minimal detection rate is curse of dimensionality also known as Hughes Phenomenon (Hughes, 1968) and imbalanced class distribution (Drummond et al., 2003). In this work a sort of class balance of the dataset is achieved by SMOTE (Chawla et al., 2002) based oversampling of the minority class and random undersampling of the majority class instances. For improving the accuracy of the model in discriminating between legitimate and illegitimate data a lot of preprocessing techniques like dimension reduction (DR), feature selection (FS), normalisation, scaling, and relabelling have been used. In Davis and Clark (2011) have surveyed the preprocessing techniques used in case of network ID, and concluded that most prominently used preprocessing technique to enhance the detection coverage and to alleviate the problems of the curse of dimensionality is DR. DR methods transform the D dimensional vector x into the d dimensional vector y with $d << D$ without sacrificing much of the information (Lee and Verleysen, 2007). Off all the DR methods applied to reduce the dimensions of the dataset principal component analysis (PCA) (Hotelling, 1933) has been most popular. PCA performs linear transformations of high dimensional (HD) onto a set of low dimensional orthogonal vectors called principal components with the aim of maximising the variance.

The world of DR has come a long way forward after the inception of linear PCA with the boundlessness of nonlinear DR techniques like EE (Carreira-Perpinán, 2010), CCA (Demartines and Hérault, 1997), ISOMAP (Balasubramanian and Schwartz, 2002), LLE (Roweis and Saul, 2000) and SNE (Hinton and Roweis, 2002), etc. have appeared on the horizon. These NLDR have proven to be far better than linear techniques in retaining the inherent information of the dataset when transformed on to lower dimensions. Motivated by the positive impact of the DR on the accuracy of the model and the ability of NLDR to effectively transform the data into lower dimensions without sacrificing much of the inherent information, this work takes a recent NLDR technique SNE to reduce the dimensions of the dataset. SNE is an unsupervised NLDR technique that obtains the ideal embedding for data items in the HD space into the LD embedding space, by minimising the Kullback-Leibler divergence between the conditional probability of the two objects being neighbors in HD space and the conditional probability of the same two points being neighbors in the low dimensional space. Fractal-based inherent dimension estimation was used to predict the lowest possible number of dimensions for which the approximation of the dataset in LD space is reasonable. On the reduced dataset a radial basis function support vector machine (SVM) that better suits the nonlinear data was implemented over five different

$Gamma$ values. To show the superiority of the proposed model a comparison of SVM on the RAW and PCA reduced data was performed. Experimental results show that proposed system has better accuracy for all the attack groups as well as the normal data. Moreover, a comparison of the proposed system with random forest which is equivocally accepted as the best classifier by the research community for multiclass classification problems has proven that the proposed model with SNE and nonlinear SVM proves to be better, which in itself is a big lead. As the novelty of the work is considered, this is the first attempt to apply any NLDR technique for network ID, as well as to implement SNE on big data (having both volume and variety) and also the pioneer to study the impact of SNE which actually is a visualisation technique on classification accuracy.

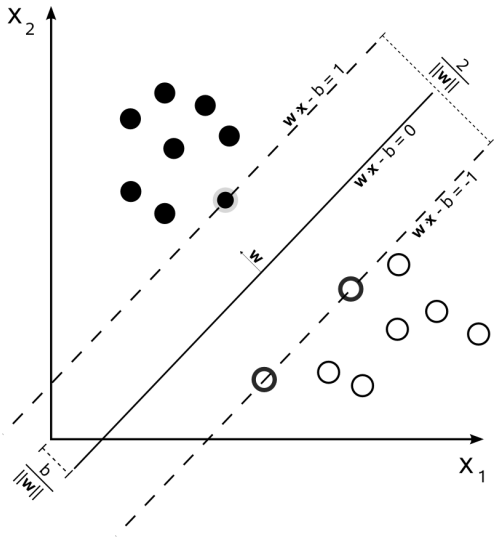
The rest of the paper is organised as follows. The motivation behind taking up this work and a brief review of the related works is given in Section 2, in Section 3 the discussion about the methods and materials used in this work are presented. Section 4 discusses the experimental methodology of the work. Results and discussions are given in Section 5, finally, Section 6 draws the conclusion.

2 Motivation and literature review

A peculiar problem in the field of network ID that has interested the research community is the low detection rate for U2R and R2L attacks. The reasons for this lower detection rate are many with imbalanced class distribution and curse of dimensionality being few. There is hardly anything that can be done to the base ML algorithm to enhance the detection rate. So any improvement that should be done should be focused on preprocessing the data before it is forwarded to the appropriate ML technique. Driven by this idea a lot of research has been carried out and appreciable results have been attained. In Tesfahun and Bhaskari (2013), Hamid et al. (2016) and Qazi and Raza (2012) researchers have used SMOTE to reduce the class imbalance and in Giacinto et al. (2008) and Michailidis et al. (2008) have implemented random undersampling on the balanced data better results have been reported. As for alleviating the curse of dimensionality is concerned FS and feature extraction have been effectively used. The area of the FS is a hot research avenue for the ML practitioners. For the purpose of selecting a subset of features in Chebrolu et al. (2005) authors have implemented Markow blanket to reduce the features to 17, in Li et al. (2009) have used wrapper-based FS to select a subset of features. In Özyer et al. (2007), Shon and Moon (2007) and Shon et al. (2006) have implemented genetic algorithms either in filter or wrapper mode to select a possibly best subset of features. As for feature extraction is concerned the only DR method that has been applied is PCA (Liu et al., 2007; Wang et al., 2004). The literature review suggests that the most common classifier in the case of ID has been SVM as in the works (Peddabachigari et al., 2007; Lin et al., 2012; Feng et al., 2014) and the most popular dataset is KDD99 dataset as in works (Kim et al., 2012; Eesa et al., 2015). There has been no reported work till date that has implemented any NLDR in ID, but NLDR has been used successfully in other classification problems (Li et al., 2015; Jamieson et al., 2010; Shekhar et al., 2014; Dupont and Ravet, 2013). All this motivated us to take up this work wherein we blend most of the pre-processing techniques. In order to balance the dataset we have applied random undersampling and SMOTE-based oversampling, to effectively implement NLDR technique we convert the nominal attributes of the dataset into numeric by arbitrary coding and this numeric

dataset is normalised so as to avoid the attributes with large feature values to dominate the attributes with lower feature values.

Figure 1 Support vector machines



3 Materials and methods

In the next few subsections, a detailed discussion about methods and materials used in this work are discussed.

3.1 Support vector machines

An SVM is a classification technique that is aimed at finding a separating plane to distinguish between two classes of instances.

From Figure 1 it can be inferred that the margin between the hyperplane and the nearest points on either side of the margin is given as $\frac{2}{||W||}$. The characteristic equation of the line separating two classes of data is $W.x - b = 0$ where W is the weight vector and b is the bias term. From a set of separating planes a decision surface that maximises the generalisation ability is selected. This is done by maximising the margin as given by the following equation.

$$\begin{aligned} \min_{W,b} \Phi(W) &= \frac{1}{2}||W||^2 \\ Subject\,to\,(W^{'} X_i + b) &\geq 1, i = 1, ...l \end{aligned} \tag{1}$$

A kernel function of the form $\psi : \Re^d \rightarrow H$ is employed to transform the data from some low dimensional space to HD space.The kernel function defines the feature space in which the hyperplane will be sought for the training data.

3.2 Radial basis kernel

Radial basis kernel is the most popular nonlinear function used to capture the similarity of the two feature vectors. Given two vectors x and y, the similarity between the two is calculated as

$$k(x, y) = exp(-\gamma||x - y||^2). \tag{2}$$

where $||x - y||^2$ is the Euclidean distance between two vectors $x\,and\,y$, and $\gamma = \frac{1}{2\sigma^2}$, σ is the only free parameter.

3.3 Random forest

Random forest (Liaw and Wiener, 2002) is a multi-faceted ensemble learning technique of decision trees that can be applied for both classification and regression. In classification, the maximum vote policy is followed whereas in regression average of the output of different trees is followed. Random forests have the inherent capability to deal better with missing data and works well with the imbalanced class dataset and also have the ability to deal with the unlabeled data.

3.4 Principal component analysis

PCA is a straightforward linear projection of data onto the principal components aimed at maximising the amount of variance. PCA model is based on the assumption that the observed variables, gathered in a random vector $y = [y_1, y_2, y_3....y_D]^{'}$ are result of some linear transformation W of P unknown latent variables, written as $x = [x_1, x_2, x_3...x_p]^{'}$ with the columns of W being orthogonal meaning $W^{'}W = I_p$ where I is an identity matrix. For a given dataset PCA finds an orthonormal basis that minimises reconstruction error

$$Error_{PCA} = \sum_{i=1}^{n} ||x_i - x_i^{'}||^2 \tag{3}$$

where $x_i^{'}$ is the transpose of vector x_i. Directed towards achieving better insight for a matrix M with n rows and m columns, a PCA seeks to uncover a set of d principal components with d being much smaller than m that preserve the maximum variance of the data matrix M. Towards reducing the dimensionality of the dataset a matrix M' of size $n \times d$ can be obtained from the matrix M by using the formula $M' = M \times P$ where P is an $n \times d$ matrix consisting of d principal components as its columns.

3.5 Stochastic neighbour embedding

Stochastic neighbour embedding is a probabilistic NLDR technique that is aimed at embedding the HD data vectors in lower dimensional spaces. SNE aims to minimise the Kullback-Leibler divergence between the similarity of the two data points in high-dimensional object space X and similarity between the same points in the lower dimensional embedding space Y. Given the asymmetric Euclidean distances matrix of

the n data points SNE kick-offs by converting the HD Euclidean distances between the data points by conditional probabilities that represent the similarities between the two data points. The similarity between the two data points in the x_i and x_j is given by the conditional probability $p_{i|j}$ that data point x_i will take data point x_j as its neighbor given by the equation

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k \neq i} exp(||x_i - x_k||^2/2\sigma^2)} \tag{4}$$

where σ_i is the only free parameter representing the variance of the Gaussian centered on the data point x_i. The conditional probability $P_{i|j}$ is inversely proportional to the Euclidean distance between the two data points. Greater the Euclidean distance between two points x_i and x_j means higher the difference between the data points x_i and x_j and lesser would be the conditional probability $P_{i|j}$, lesser the Euclidean distance between the two points, higher would be the value of conditional probability.

The data points y_i and y_j represent the embedding of the two data points x_i and x_j. The conditional probability of the same two points in the lower dimensional embedding space is given by the conditional probability $q_{i|j}$ that the data point y_i will take up the data point y_j as its neighbor as given by the following equation

$$q_{j|i} = \frac{exp(-||y_i - y_j||^2/2\sigma^2)}{\sum_{k \neq i} exp(||y_i - y_k||^2/2\sigma^2)} \tag{5}$$

By setting the variance of the Gaussian centered on the object y_i in the embedding space equal to $\frac{1}{\sqrt{2}}$, above equation can be written as

$$q_{j|i} = \frac{exp(-||y_i - y_j||^2)}{\sum_{k \neq i} exp(||y_i - y_k||^2)} \tag{6}$$

If the embedding spaces map the feature space correctly for all the data points the conditional probabilities $p_{i|j}$ and $q_{i|j}$ will be equal. The confidence with which $q_{j|i}$ models the $p_{j|i}$ is given by the Kullback-Leibler divergence

$$KLD(P_i||Qi) = \sum_j p_{j|i}\ log\frac{p_{j|i}}{q_{j|i}} \tag{7}$$

The ideal embedding is given by minimising the Kullback-Leibler divergence W over all the points i and j as given by the equation

$$W = \sum_i KLD\ (P_i||Q_i) \tag{8}$$

The cost function given in the above equation is minimised by the gradient given in the equation below

$$\frac{\delta W}{\delta y_i} = 2\sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j) \tag{9}$$

3.6 KDD99 intrusion dataset

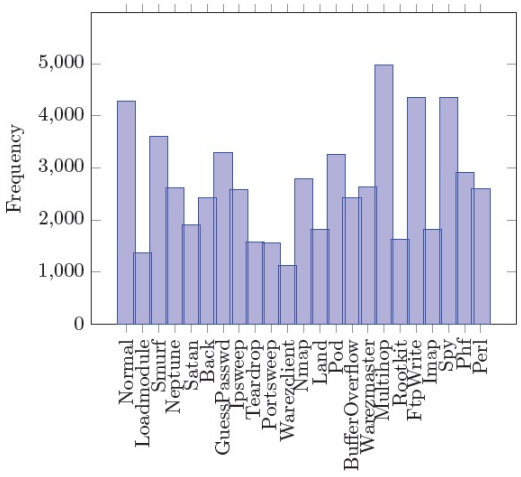
KDD99 and UNSW-NB15 datasets are mostly used and widely accepted benchmark datasets for network ID. Of course, there are many other datasets like DDoS dataset (Moore and Shannon, 2003), UNM dataset, ADFA Linux dataset (Creech and Hu, 2013) to name a few, that have surfaced up in last few years but have attained very less acceptance among the diaspora of network security engineers till now. As for the DOS dataset is concerned it covers only one group of attacks, i.e., DDoS, being specialised the dataset cannot be used for generalised IDS. While as ADFA dataset is suitable for HOST-based systems only and has the traffic captured on the single host it does not take into consideration the distributed attacks making it infeasible. KDD99 and UNSW-NB15 datasets alone have the proper mix of all the attacks be it network-based or host-based, and moreover they cover all the attacks that get surfaced up form the user trying to locate the victim (probe) for gaining initial access to the victim (R2L) and gaining superuser privileges (U2R). There are only a few works that have used the other mentioned datasets and as pointed out correctly by authors in Lin et al. (2015) and Liao et al. (2013) KDD99 has been the most popular of them all, so using these two datasets in our work will lead to better comparison of the model with other works. All these problems pertinent in other datasets make the usage of KDD99 and UNSW-NB15 indispensable. The KDD99 dataset is mixed in nature with some attributes being numeric and some being nominal. In total 41 attributes are used to represent a connection, and 42^{nd} attribute signifies whether the connection is normal or an attack. A total 23 attacks spanning across four groups, user to root, remote to local, DOS, and probe are present in the dataset. The frequency of the different attacks and also the normal data varies for each group as shown in Figure 2. Most predominant attack groups in the dataset are DoS and Probe. There are lesser records pertaining to U2R and R2L attacks. In addition to four attack groups, there is a group for normal data as well. So, Let $\mathbf{X} = (x_1, ..., x_n)^T$ be the $n \times m$ data matrix. The number n represents the number of samples used in the dataset, and m the number of attributes. We have in our case n = 61,906 and m = 41. This dataset represents a big data challenge in a high dimension context.

3.7 UNSW-NB15 dataset

UNSW-NB15 (Moustafa and Slay, 2015) dataset was created in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) by the IXIA perfect storm tool (Hughes, 1968). This dataset contains a hybrid of normal activities and attack behaviours. Tcpdump tool is used to capture 100 GB of the raw trace. Twelve algorithms and tools such as Argus, Bro-IDS are used to generate UNSW-NB15. In total the dataset is comprised of 49 features which also includes the class label. These attributes determine the features of the connections. The attributes are mixed in nature with some being nominal, some being numeric and some being taking time-stamp values. This dataset contains a total of 2,540,044 labelled instances categorised as being either normal connections or attack connections. A total of nine attack categories namely fuzzers, reconnaissance, shellcode, analysis, backdoors, DoS, exploits, generic and worms are present in the dataset in addition to the instances representing the normal connection. This dataset captured and made online in 2015 only has a vast amount of attacks as compared to KDD99. So as matter of fact this dataset can be used for ID with

confidence. As was the case with KDD99 some features of UNSW-NB15 were nominal like protocol, state, etc. In order to apply the mathematical models on this data, we replaced the nominal features by their equivalent values. In total, we have taken 1 lakh instances comprising of both normal and malicious connections.

Figure 2 Analysis of different classes (see online version for colours)



3.8 Estimating intrinsic dimensionality

The basic assumption of the work is that, even though data points are points in $\mathbb{R}^m$ which in this case is 41, there exists a p-dimensional manifold $\mathcal{M}$ with $p << m$ that can satisfyingly represent all the inherent information of the dataset. The lowest possible value of p for which the approximation of $\mathbf{X}$ by $\mathcal{M}$ is reasonable is called the ID of $\mathbf{X}$ in $\mathbb{R}^m$. For estimating ID of our KDD datasets, we in this work applied a geometric procedure that estimates the equivalent notion of fractal dimension (Camastra and Vinciarelli, 2002). The estimated intrinsic dimensionality of both KDD and UNSW-NB15 dataset using fractal dimension was found out to be p = 3. So we retained three dimensions for both PCA and SNE.

4 Methodology

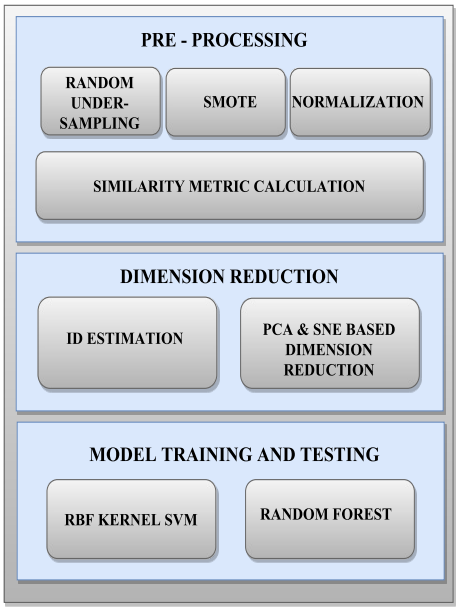
As shown in Figure 3, the proposed methodology is split into three parts, i.e., data preprocessing, DR and model training and testing. The following subsections discuss in detail all the three phases.

4.1 Data pre-processing

The work starts with selecting a subset of full KDD dataset, as using the full KDD dataset is practically not possible. For the randomly drawn subset, a sort of balance

among the instances of different classes is attained by random under-sampling of the majority class instances and systematic oversampling of the minority classes by using SMOTE. As the data is somewhat balanced, the nominal attributes are converted into numeric attributes by the arbitrary coding of different nominal values. This numeric dataset with the nominal class variable is subjected to normalisation to the range [0–1] so as to avoid the features with very high values from dominating the features with small values.

Figure 3 Block diagram of proposed system (see online version for colours)



4.2 Dimension reduction

Before exposing the data to DR we study the intrinsic dimensionality of the dataset using the fractal analysis. Only when we have a clear idea about how many features are actually enough for classification, the data can be projected effectively. From the fractal analysis, we could find out that at minimum three features are essential for classifying this dataset. We then compute an asymmetric Euclidean distance matrix for the data points. On this matrix only SNE is implemented and the data is transformed, from the transformed data first three dimensions are retained. In addition to implementing SNE we also applied PCA on balanced and normalised data. In this case, also we retained three dimensions.

4.3 Model training and testing

Once we have the reduced data, i.e., data is transformed to $n \times 4$ form (where the first three attributes represent projected data and the last attribute designates class label) we proceed by training and testing SVM on all the versions of the dataset, i.e., RAW,

PCA and SNE. For each dataset, we train the nonlinear SVM based on radial bias function using five different $gamma$ values $0.01, 0.1, 1, 10, 100$. For each value of $gamma$, various classification measures like $correctlyclassified$, $TpRate$, $precision$ and $recall$ are taken into consideration. In addition to that, the performance of the proposed is also compared to the random forests over all the three variations of the dataset.

On the reduced datasets RBF kernel-based SVM with five different gamma values $0.01, 0.1, 1, 10, 100$ is trained and tested. On the part of exhibiting the effectiveness of the DR, the performance results of SVM on reduced data are compared with those obtained from the RAW data. In addition to comparing the results of SVM on SNE reduced dataset to those of PCA and RAW data, a comparison of SVM with random forests has also been carried out.

As for the novelty of this work is considered this is the first attempt to apply any NLDR technique for network ID. Over the decades a set of linear DR like PCA has been applied for reducing the dimensionality of KDD dataset. Nonlinear DR techniques attain very high data reduction whilst retaining the maximum information of the data. This better aversion of the curse of dimensionality leads to the better detection for minimal classes of attacks, i.e., U2R and R2L which have been the cause of menace for ID professionals and researchers. Since the dataset is projected to only three dimensions making it very feasible for the application in any problem domain suffering from the curse of dimensionality.

5 Results and discussions

In the following subsections, we present the results of the proposed system. The results are provided along four dimensions, first, in the Subsection 5.1, a 2D scatter plot of SNE reduced 3D data. In Subsection 5.2 results of the experiments on SVM are provided. Subsection 5.3 presents a comparison of the proposed work with other prominent works in the field is given. In Subsection 5.4 a comparison of the proposed system based on SNE with SVM and Random forests is given.

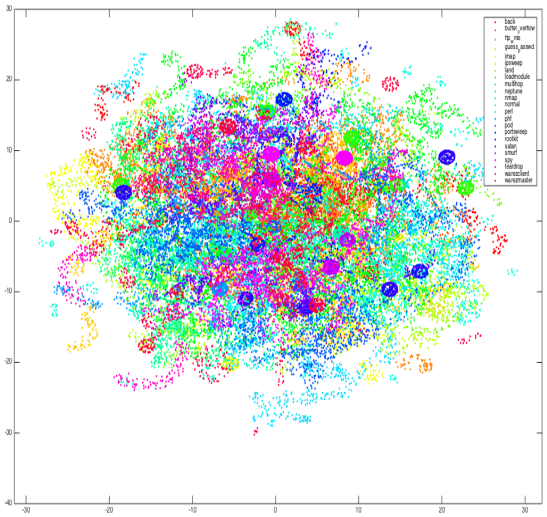
5.1 KDD99 results

In the next few subsections, we present the results of the proposed model on the KDD99 dataset.

5.1.1 Scatter plot

A scatter plot of the SNE reduced data on the 2D plane is given in Figure 4. As it is pretty clear from the figure that t-SNE helps in the effective clusterisation of the data. From the scatter plot it is apparent that the reduced data results in well-separated clusters when projected on the 2D plane. The records belonging to different classes are clustered more or less very close to each other and very farther from the other class data.

Figure 4 Block diagram of proposed system (see online version for colours)



5.1.2 SVM performance metrics

Given in Figures 5 to 9 report various performance metrics of SVM on different values of gamma for The KDD99 dataset. Each of the figures represents the particular performance metric of SVM on all the three variations of the dataset, i.e., RAW data, SNE reduced and PCA reduced. The performance metrics across which different works are compared are $detectionrate$, $TP-rate$, $FP-rate$, $precision$, $recall$ and $ROCarea$.

In Figure 5, we present a graph plotting $detectionrate$ of SVM on all the three datasets against all five different values of $gamma$ for the KDD99 dataset. Starting with the value $gamma$ = 0.01 the $detectionrate$ of SVM is around 96% for the RAW data and is 84% for the PCA reduced data and is lowest close to 80% for the SNE reduced data. As the value of $gamma$ changes from 0.01 to 0.1 the performance of SVM on both the reduced dataset increases very appreciably, from there onwards the performance of SVM increases moderately but at the same time the $detectionrate$ keeps diminishing continuously at the value of $gamma$ = 100 the $detectionrate$ of SVM is highest on the SNE reduced data followed by the PCA reduced data and is lowest on the raw data.

In Figure 6, we present a graph plotting $truepositiverate$ of SVM on all the three datasets against all five different values of $gamma$ for KDD99 dataset. As in case of $detection\ rate$ at the lowest value of $gamma$ SVM reports best $truepositiverate$ for RAW data. But as the value of $gamma$ increases, the $TPrate$ of SVM starts to diminish continuously on the RAW data, and that on the both the reduced datasets is on a continuous increase. At the highest value of the gamma that we have taken in the course of this work, SVM has the highest $TPrate$ close to one on SNE reduced data,

followed by the PCA reduced data, and has the lowest $TPrate$ of 0.68 on the RAW data.

Figure 5 Detection rate (see online version for colours)

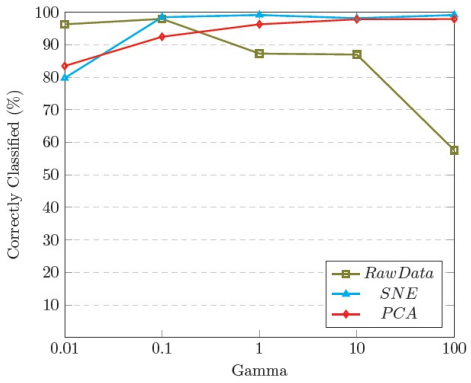


Figure 6 TP rate (see online version for colours)

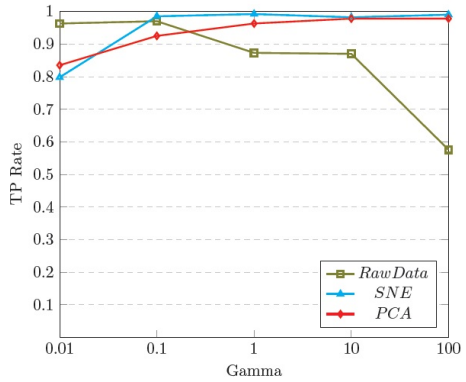


Figure 7 Precision (see online version for colours)

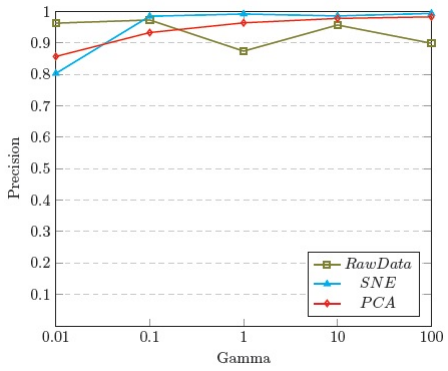


Figure 8 Recall (see online version for colours)

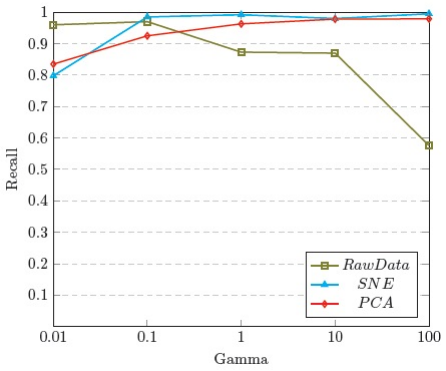
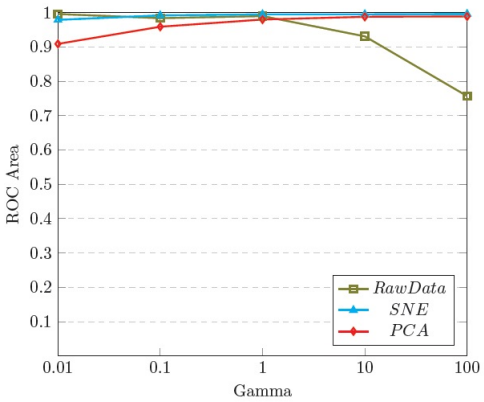


Figure 9 ROC area (see online version for colours)



In Figure 7 we present a graph plotting $precision$ of SVM on all the three datasets against all five different values of $gamma$ for KDD99 dataset. Beginning with the value of $gamma$ = 0.01 SVM reports lowest $precision$ on SNE reduced dataset, and has the highest $precision$ on the RAW data. At the $Gamma$ value of 0.1 both SVM reports the similar precision value on both RAW and SNE reduced data. Likewise, the $precision$ also improved for PCA reduced data but is still less than that on two other datasets. At the highest value of $gamma = 100$, the SVM has highest $precision$ on the SNE reduced data and the lowest on the RAW data. There is only a slight difference between the $precision$ of SVM on both the reduced datasets.

In Figure 8 we present a graph plotting $recall$ of SVM on all the three datasets against all five different values of $gamma$ for the KDD99 dataset. At the $gamma$ = 0.01, SVM has better performance for the RAW data when compared to PCA and SNE reduced datasets. As the value of $gamma$ increases to 0.1 the performances of SVM diminishes on the RAW data and increases on both the reduced dataset. At the value $gamma$ = 100 SVM reports lowest $recall$ for RAW data and highest $recall$ on SNE reduced data. In the course of experiments on both reduced dataset $recall$ increases with the increase in the $gamma$ value.

In Figure 9 we present a graph plotting ROC of SVM on all the three datasets against all five different values of $gamma$ for the KDD99 dataset. At the lowest value of $gamma$ SVM starts with the ROC value close to one for both RAW data and is very less on the SNE reduced data. As the value of $gamma$ changes from 0.01 to one, ROC is somewhat steady on both the RAW and PCA reduced data but is on a continuous increase on SNE reduced data. At the highest value of $gamma$, i.e., 100, SVM reports highest ROC close to one on SNE reduced data, followed by that on the PCA reduced data, and is lowest around 0.76 on the RAW data.

5.1.3 Comparison of related works

In given in Table 1 a comparison of the proposed system with various other works is given. From the data given in the table, it is clear that the proposed system has better detection rate than all the other works for all the attack groups as well as the normal data. For Dos attacks alone the model proposed by in Horng et al. (2011) performs better than our proposed model. But the difference is very small. For all other, the attack groups and normal data our proposed model reports much higher detection rates.

Table 1 Comparison with related works

Model	*Normal*	*DOS*	*Probe*	*R2L*	*U2R*
Mukkamala et al. (2002)	97.22	97.87	77.19	86.76	19.14
Hernández-Pereira et al. (2009)	98.73	92.63	52.56	0.00	0.00
Toosi and Kahani (2007)	98.20	99.50	84.10	31.50	14.12
Lin et al. (2015)	95.98	82.85	96.59	78.95	61.54
Tsai and Lin (2010)	96.12	83.12	96.59	60.00	78.26
Peddabachigari et al. (2007)	99.64	99.78	61.72	28.06	40.00
Proposed system	*99.85*	98.89	*98.89*	*96.51*	*86.80*

5.2 UNSW results

In the next few subsections, we present the results of the proposed model on UNSW dataset.

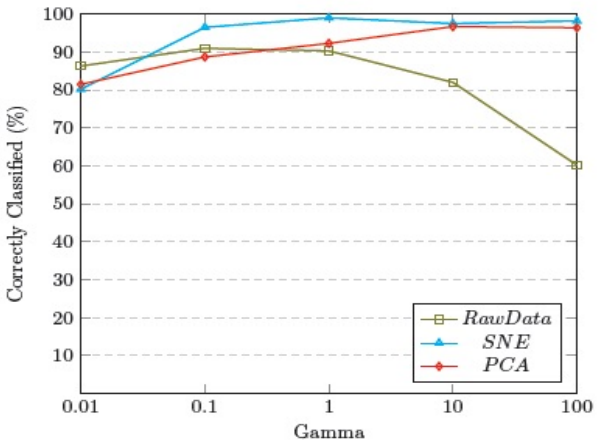
5.2.1 SVM performance metrics

Given in Figures 10 to 14 report various performance metrics of SVM on different values of gamma for the UNSW-NB15 dataset. Each of the figures represents the particular performance metric of SVM on all the three variations of datasets, i.e., RAW data, SNE reduced and PCA reduced. The performance metrics across which different works are compared are $detectionrate$, $TP-rate$, $FP-rate$, $precision$, $recall$ and $ROCarea$.

In Figure 10, we present a graph plotting $detectionrate$ of SVM on all the three datasets against all five different values of $gamma$. As can be seen on from the figure that the proposed model with SNE reduced data reports the better detection rate. We

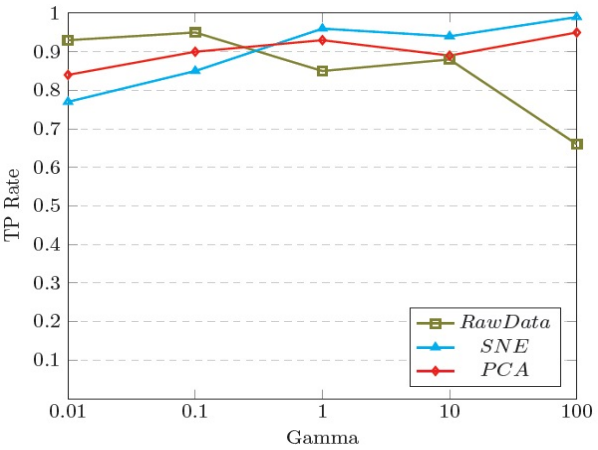
have taken the detection rate for all the attack groups separately and plotted an average of all the values.

Figure 10 Detection rate (see online version for colours)



In Figure 11, we present a graph plotting $truepositiverate$ of SVM on all the three variations of the dataset against all five different values of $gamma$. As it is perfectly evident from the figure that proposed model with SNE has higher $truepositiverate$ than all the other models. The higher $truepositiverate$ is driven by the fact that the proposed model is able to preserve the maximum amount of information in the least number of dimensions.

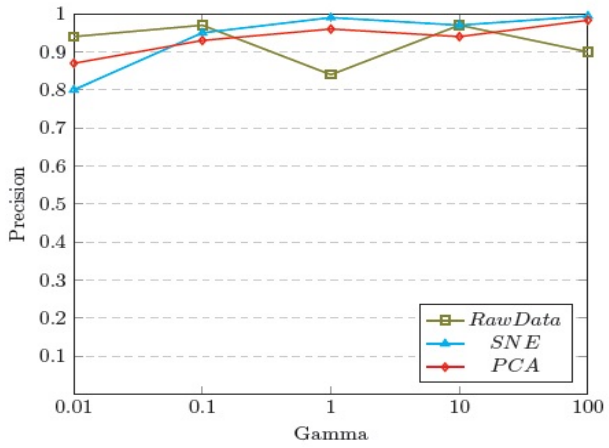
Figure 11 TP rate (see online version for colours)



In Figure 12 we present a graph plotting $precision$ of SVM on all the three datasets against all five different values of $gamma$ for the UNSW-NB15 dataset. Starting with the value of $gamma = 0.01$ SVM reports lowest $precision$ on SNE reduced dataset,

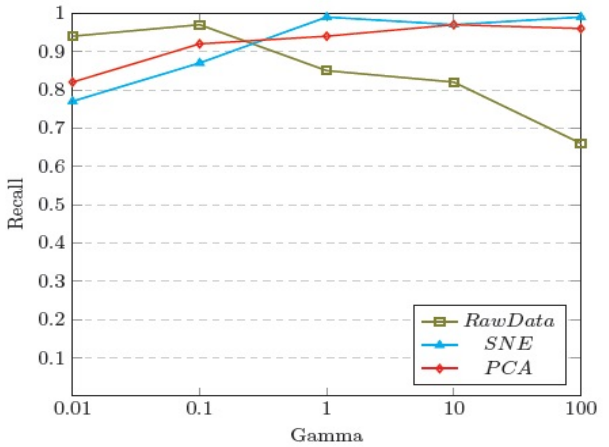
and has the highest $precision$ on the RAW data. With the increase in $gamma$ value, the $precision$ of SVM increases for SNE reduced data. Even though for PCA reduced data also $precision$ increases with the increase in the $gamma$ value it is still lesser than SNE reduced data.

Figure 12 Precision (see online version for colours)



In Figure 13 we present a graph plotting $recall$ of SVM on all the three datasets against all five different values of $gamma$ on the UNSW-NB15 dataset. As can be seen from the figure the proposed model has higher recall than all the other models on most of the $gamma$ values. With the growing $gamma$ values, the $recall$ of the proposed model with SNE also increases appreciably.

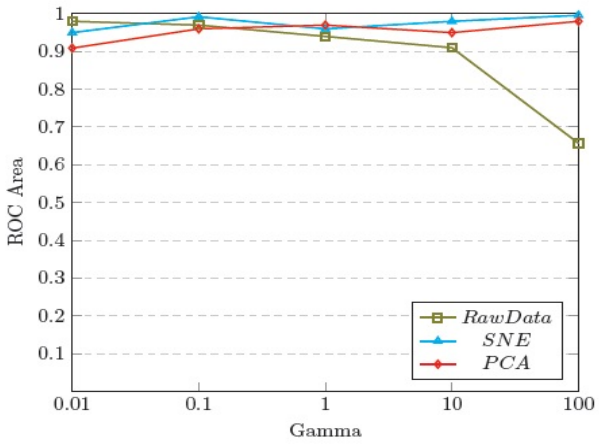
Figure 13 Recall (see online version for colours)



In Figure 14 we present a graph plotting ROC of SVM on all the three datasets against all five different values of $gamma$ for the UNSW-NB15 dataset. As can be seen from

Figure 14 SVM reports better ROC on the proposed model of SNE reduced data than all the variations of the data.

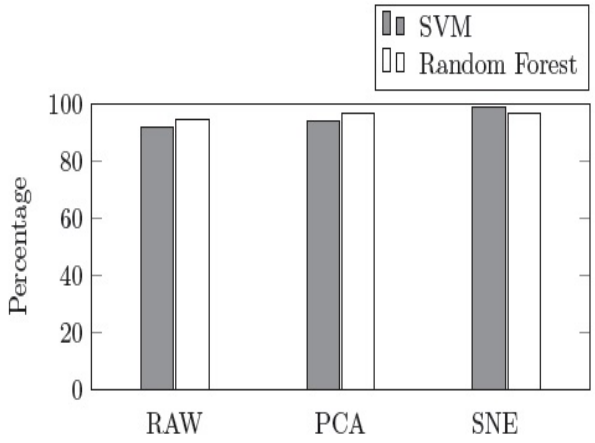
Figure 14 ROC area (see online version for colours)



5.3 Comparison with random forests for KDD dataset

In Figure 15, a comparison of SVM and random forests is drawn. On the multi-class classification problems, Random Forest an ensemble of decision tree classifier more often gives better results than the single classifier. Here in this work, we made use of random forests to check the effectiveness of SNE in improving the detection rate of nonlinear SVM for network ID. On both raw data and PCA reduced datasets Random Forests report better detection rate than SVM. But in the case of SNE reduced data, the proposed system using SVM with RBF outperforms the random forests, which in itself is a big lead.

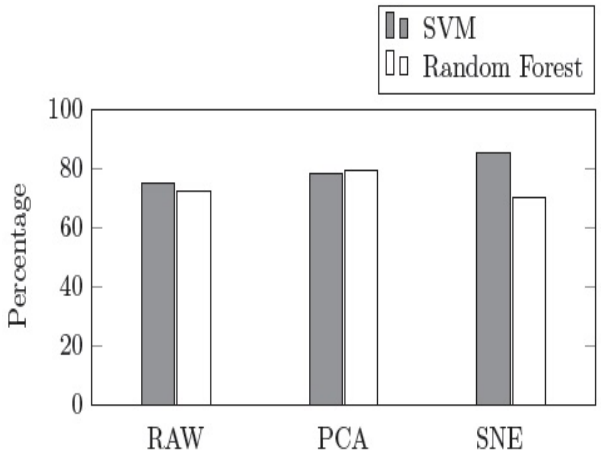
Figure 15 Detection rate on variations of KDD99 dataset



5.4 Comparison with random forests for UNSW

In Figure 16, a comparison of SVM and random forests on the UNSW-NB15 dataset is drawn. As already mentioned above that on the multi-class classification problems, random forest an ensemble of decision tree classifier more often gives better results than the single classifier. As done for the KDD99 dataset we performed the comparison of the proposed model on the UNSW-NB15 dataset with random forests. As can be seen from the figure the proposed model outperforms the random forest.

Figure 16 Detection rate on variations of UNSW-NB15 dataset



5.5 Discussion

A comparison of the results of SVM on RAW, PCA reduced and SNE reduced data support the claim that both KDD99 and UNSW-NB15 datasets suffered the curse of dimensionality, which results in the hampering the detection coverage of the model. All through the experiments, SVM performs better on SNE reduced data as compared to RAW and PCA reduced data over all the performance metrics $TPRate$, $detectionrate$ and ROC, etc. At the beginning, in all the experiments SVM performs better on RAW data but as the $Gamma$ value increases the performance of SVM drops on the RAW data and increases on the SNE and PCA reduced data. At the highest value of $gamma$, the SVM performs best on SNE reduced data, followed by PCA and RAW data. A comparison of the proposed system with the random forest that actually gives the best results on the multiclass classification problem shows that our proposed system has better detection rate than the RF which in itself is an improved edge of the work. Moreover, we compared the proposed model to the already existing works in the field proved that the proposed model has better detection rate for all attack groups than all of the other techniques. The other benefit of this work is that proposed work was able to obtain the better performance than other works at the same time using only three transformed attributes from a total of 41 attributes that the model started with which is a big gain over the amount of resource required to hold the data.

6 Conclusions

This work as the first of its kind applies SNE a probabilistic NLDR technique to reduce the dimensionality of ID dataset. This is the first attempt to apply SNE on big data and also the first attempt to experiment the effects of the data visualisation techniques on classification accuracy. The main motivation of this work was lower detection rates of IDS models for minimal attack groups, which result from the curse of dimensionality. The intended work projects the data onto three dimensions as predicted by inherent dimensionality estimation. On the reduced dataset a nonlinear SVM based on radial basis function was tested over five different values of $gamma$. The experimental results be it the SVM performance metrics, comparison with random forests and the compassion with the prominent works in the field show that the proposed model has better detection rate overall the attack groups as well as on the normal data for both the datasets. This improved detection rate of the proposed system comes with complex DR process of the dataset. In future, we will try to replace the Euclidean distance metric with GOWER index metric that will be better suited for mixed data. We hope that by changing the distance metric we can have very well separated clusters in the dataset, which in turn will have a positive impact on classification performance metrics.

References

Axelsson, S. (2000) *Intrusion Detection Systems: a Survey and Taxonomy*, Technical Report.

Balasubramanian, M. and Schwartz, E.L. (2002) 'The isomap algorithm and topological stability', *Science*, Vol. 295, No. 5552, p.7.

Camastra, F. and Vinciarelli, A. (2002) 'Estimating the intrinsic dimension of data with a fractal-based method', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 10, pp.1404–1407.

Carreira-Perpinán, M.A. (2010) 'The elastic embedding algorithm for dimensionality reduction', in *ICML*, Vol. 10, pp.167–174.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) 'SMOTE: synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, Vol. 16, pp.321–357.

Chebrolu, S., Abraham, A. and Thomas, J.P. (2005) 'Feature deduction and ensemble design of intrusion detection systems', *Computers and Security*, Vol. 24, No. 4, pp.295–307.

Creech, G. and Hu, J. (2013) 'Generation of a new IDS test dataset: Time to retire the KDD collection', in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, IEEE, pp.4487–4492.

Davis, J.J. and Clark, A.J. (2011) 'Data preprocessing for anomaly based network intrusion detection: a review', *Computers and Security*, Vol. 30, No. 6, pp.353–375.

Demartines, P. and Hérault, J. (1997) 'Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of datasets', *IEEE Transactions on Neural Networks*, Vol. 8, No. 1, pp.148–154.

Drummond, C., Holte, R.C. et al. (2003) 'C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling', *Workshop on Learning from Imbalanced Datasets II*, Citeseer, Vol. 11.

Dupont, S. and Ravet, T. (2013) 'Improved audio classification using a novel nonlinear dimensionality reduction ensemble approach', in *ISMIR*, Citeseer, pp.287–292.

Eesa, A.S., Orman, Z. and Brifcani, A.M.A. (2015) 'A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems', *Expert Systems with Applications*, Vol. 42, No. 5, pp.2670–2679.

Feng, W., Zhang, Q, Hu, G. and Huang, J.X. (2014) 'Mining network data for intrusion detection through combining SVMs with ant colony networks', *Future Generation Computer Systems*, Vol. 37, pp.127–140.

Giacinto, G., Perdisci, R., Del Rio, M. and Roli, F. (2008) 'Intrusion detection in computer networks by a modular ensemble of one-class classifiers', *Information Fusion*, Vol. 9, No. 1, pp.69–82.

Hamid, Y., Sugumaran, M. and Balasaraswathi, V.R. (2016) 'IDS using machine learning-current state of art and future directions', *British Journal of Applied Science and Technology*, Vol. 15, No. 3, pp.1–22.

Hamid, Y., Hamid, Y., Sugumaran, M. and Journaux, L. (2016) 'A fusion of feature extraction and feature selection technique for network intrusion detection', *International Journal of Security and its Apllications*, Vol. 10, No. 8, pp.151–158.

Hernández-Pereira, E. Suárez-Romero, J.A., Fontenla-Romero, O. and Alonso-Betanzos, A. (2009) 'Conversion methods for symbolic features: a comparison applied to an intrusion detection problem', *Expert Systems with Applications*, Vol. 36, No. 7, pp.10612–10617.

Hinton, G.E. and Roweis, S.T. (2002) 'Stochastic neighbor embedding', in *Advances in Neural Information Processing Systems*, pp.833–840.

Horng, S-J., Su, M-Y., Chen, Y-H., Kao, T-W., Chen, R-J., Lai, J-L. and Perkasa, C.D. (2011) 'A novel intrusion detection system based on hierarchical clustering and support vector machines', *Expert Systems with Applications*, Vol. 38, No. 1, pp.306–313.

Hotelling, H. (1933) 'Analysis of a complex of statistical variables into principal components', *Journal of Educational Psychology*, Vol. 24, No. 6, p.417.

Hughes, G. (1968) 'On the mean accuracy of statistical pattern recognizers', *IEEE Transactions on Information Theory*, Vol. 14, No. 1, pp.55–63.

Jamieson, A.R., Giger, M.L., Drukker, K., Li, H., Yuan, Y. and Bhooshan, N. (2010) 'Exploring nonlinear feature space dimension reduction and data representation in breast CADx with Laplacian eigenmaps and t-SNE', *Medical Physics*, Vol. 37, No. 1, pp.339–351.

Kim, G., Lee, S. and Kim, S. (2012) 'A novel hybrid intrusion detection method integrating anomaly detection with misuse detection', *Expert Systems with Applications*, Vol. 41, No. 4, pp.1690–1700.

Lee, J.A. and Verleysen, M. (2007) *Nonlinear Dimensionality Reduction*, Springer Science and Business Media, Berlin, Heidelberg.

Li, Y., Wang, J-L., Tian, Z-H., Lu, T-B. and Young, C. (2009) 'Building lightweight intrusion detection system using wrapper-based feature selection mechanisms', *Computers and Security*, Vol. 28, No. 6, pp.466–475.

Li, Y., Wang, Yu., Zi, Y. and Zhang, M. (2015) 'An enhanced data visualization method for diesel engine malfunction classification using multi-sensor signals', *Sensors*, Vol. 15, No. 10, pp.26675–26693.

Liao, H-J., Lin, C-H.R, Lin, Y-C. and Tung, K-Y. (2013) 'Intrusion detection system: a comprehensive review', *Journal of Network and Computer Applications*, Vol. 36, No. 1, pp.16–24.

Liaw, A. and Wiener, M. (2002) 'Classification and regression by randomForest', *R news*, Vol. 2, No. 3, pp.18–22.

Lin, S-W., Ying, K-C., Lee, C-Y. and Lee, Z-J. (2012) 'An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection', *Applied Soft Computing*, Vol. 12, No. 10, pp.3285–3290.

Lin, W-C., Ke, S-W. and Tsai, C-F. (2015) 'CANN: An intrusion detection system based on combining cluster centers and nearest neighbors', *Knowledge-based Systems*, Vol. 78, pp.13–21.

Liu, G., Yi, Z. and Yang, S. (2007) 'A hierarchical intrusion detection model based on the PCA neural networks', *Neurocomputing*, Vol. 70, No. 7, pp.1561–1568.

Michailidis, E., Katsikas, S.K. and Georgopoulos, E. (2008) 'Intrusion detection using evolutionary neural networks', *Panhellenic Conference on Informatics*, PCI'08, pp.8–12.

Moore, D. and Shannon, C. (2003) *SCO Offline from Denial of Service Attack* [online] http://www.caida.org/analysis/security (accessed 26 October 2016).

Moustafa, N. and Slay, J. (2015) 'UNSW-NB15: a comprehensive dataset for network intrusion detection systems (UNSW-NB15 network dataset)', in *Military Communications and Information Systems Conference (MilCIS)*, IEEE, pp.1–6.

Mukkamala, S., Janoski, G. and Sung, A. (2002) 'Intrusion detection using neural networks and support vector machines', in *Proceedings of the 2002 International Joint Conference on Neural Networks*, IJCNN'02, IEEE, pp.1702–1707.

Özyer, T., Alhajj, R. and Barker, K. (2007) 'Intrusion detection by integrating boosting genetic fuzzy classifier and data mining criteria for rule pre-screening', *Journal of Network and Computer Applications*, Vol. 30, No. 1, pp.99–113.

Peddabachigari, S., Abraham, A., Grosan, C. and Thomas, J. (2007) 'Modeling intrusion detection system using hybrid intelligent systems', *Journal of Network and Computer Applications*, Vol. 30, No. 1, pp.114–132.

Qazi, N. and Raza, K. (2012) 'Effect of feature selection, SMOTE and under sampling on class imbalance classification', in *2012 UKSim 14th International Conference on Computer Modelling and Simulation (UKSim)*, pp.145–150.

Roweis, S.T. and Saul, L.K. (2000) 'Nonlinear dimensionality reduction by locally linear embedding', *Science*, Vol. 290, No. 5500, pp.2323–2326.

Shekhar, K., Brodin, P., Davis, M.M. and Chakraborty, A.K. (2014) 'Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE)', *Proceedings of the National Academy of Sciences*, Vol. 111, No. 1, pp.202–207.

Shon, T. and Moon, J. (2007) 'A hybrid machine learning approach to network anomaly detection', *Information Sciences*, Vol. 177, No. 18, pp.3799–3821.

Shon, T., Kovah, X. and Moon, J. (2006) 'Applying genetic algorithm for classifying anomalous TCP/IP packets', *Neurocomputing*, Vol. 69, No. 16, pp.2429–2433.

Stallings, W. (2007) *Network Security Essentials: Applications and Standards*, Pearson Education, India.

Tesfahun, A. and Bhaskari, D.L. (2013) 'Intrusion detection using random forests classifier with SMOTE and feature reduction', in *2013 International Conference on Cloud and Ubiquitous Computing and Emerging Technologies (CUBE)*, pp.127–132.

Toosi, A.N. and Kahani, M. (2007) 'A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers', *Computer Communications*, Vol. 30, No. 10, pp.2201–2212.

Tsai, C-F. and Lin, C-Y. (2010) 'A triangle area based nearest neighbors approach to intrusion detection', *Pattern Recognition*, Vol. 43, No. 1, pp.222–229.

Wang, W., Guan, X. and Zhang, X. (2004) 'A novel intrusion detection method based on principle component analysis in computer security', in *International Symposium on Neural Networks*, Springer, Berlin, Heidelberg, pp.657–662.