NETWORK ANOMALY DETECTION SYSTEM USING MACHINE LEARNING TECHNIQUE A PROPOSED MODEL

***Sujeet Raosaheb Suryawanshi[1] and Kalpana Thakre[2]**
[1]NSDL e-Governance Infrastructure Limited, Mumbai, India
[2]Department of Information Technology, Sinhgad College of Engineering, Pune, India
**Author for Correspondence*

ABSTRACT

Increase in use of information systems has increased the need to secure network access of these information systems. In order to detect attacks, security solutions such as Intrusion Detection and Prevention Systems (IDPS) are deployed. IDPS is the most basic way to protect a network. These systems follow three different approaches to detect malicious intended network intruders' viz., signature-based, rule-based/specification-based or anomaly-based detection (AD). Signature-based and rule-based systems are ineffective. Anomaly based techniques can utilize Machine Learning (ML) techniques to detect intrusive packet. This study focuses on to provide a systematic review of ML techniques that can be used for AD of Denial of Service attack detection and propose possible model for designing Network Anomaly system. ML techniques for AD are based on distance or density or statistical based algorithms, sometimes combination of more than one is used to achieve higher accuracy in terms of detection; these are called as ensemble or hybrid ML techniques.

Keywords: *Machine Learning, Anomaly Detection, Denial of Service, Clustering, Classification, Intrusion Detection System, Intrusion Prevention System, Ensemble*

INTRODUCTION

Due to increased complexity in information systems, there is an increase in threats and vulnerability. Because of constrained time to ascertain the attack situation and act to normalize and protect the system from further damage, Intrusion Detection and Prevention Systems (IDPS) are the most basic way to protect the network access from attacks. Mostly, IDPS systems follow two different approaches for detecting security breaches: signature-based or anomaly-based detection. A signature-based IDPS compares network packets against a set of defined signatures (known threats). Signature based method has limitations as it can only identify known threats, e.g. abstract, or extremely generic, signature. The anomaly-based detection technique centers on the concept of a base lining network behavior and flagging it as *anomaly* if there is any deviation to the same.

Classification of Intrusion Detection and Prevention System (IDPS)

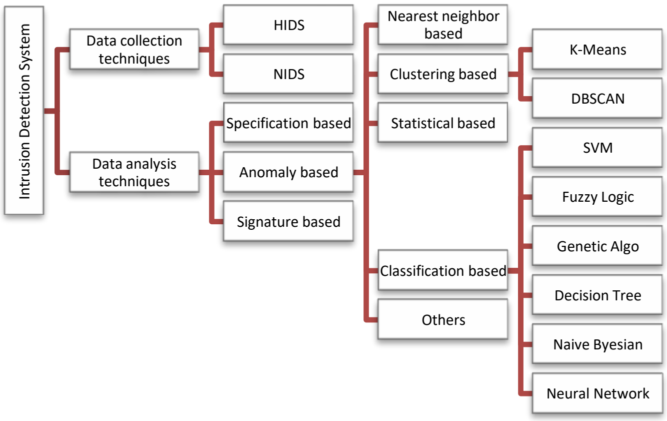

Figure 1 Classification of IDPS

From data collection mechanism per se, IDPS can be classified into Host-based and Network-based (Karami and Kheyri, 2012).

In Network-based IDPS, packets are captured from network, and analyzed; in Host-based IDPS, packets are capture through host system itself.

As per Data analysis techniques, IDPS has below mentioned methods for attack detection:

Signature Based: Known attacks signatures are configured. Packet signatures matching with configured signature is marked as probable attack.

Specification Based: Rules are configured to describe the expected operation of protocol. Any deviation from the expected operation is marked as probable attack.

Anomaly Based: Normal behavior of the target system (network and nodes) is base lined, and then a profile or normal behavior model is constructed according to it. Based on the profile a threshold is defined. Nodes are monitored; incase of any unmatched behavior, it is marked as a probable attack.

Different Aspects of Machine Learning

This section identifies and discusses the different aspects of anomaly detection.

a) Model Markup Language:

The Predictive Model Markup Language (PMML) is an XML-based language that helps to define models that can be used with different PMML compliant applications.

PMML provides applications a vendor-independent method of defining models so that proprietary issues and incompatibilities are no longer a barrier to the exchange of models between various applications. It helps to develop models once, and use with different applications.

b) Output of Anomaly Detection

An important aspect for any anomaly detection system using machine learning is the technique devised for reporting anomalies detected.

Scores: Each data instance is assigned an anomaly score depending on the degree to which that instance is assumed to be an anomaly. Cut-off or threshold technique can be used to decide on which instance to be reported as anomaly.

Labels: Data instance is assigned a label (normal or anomalous).

c) Other Aspects

Machine Learning techniques such as SVM with PCA or logistic regression with Random Forest are effective in performance and are largely researched. Effectiveness is measured in terms of reduction of false detections (False positives and false negatives).

Design Approach

We would follow the basic methodology as explained in (Gutierrez and Branch, 2012).

a) Basic Methodology of Anomaly Detection Technique

Involves selecting data, analyzing attributes, marking data for training and testing, analyzing effectiveness of the model. We would be using multiple algorithms for same set of data and compare the effectiveness.

b) Real Time Outlier Detection System Architecture

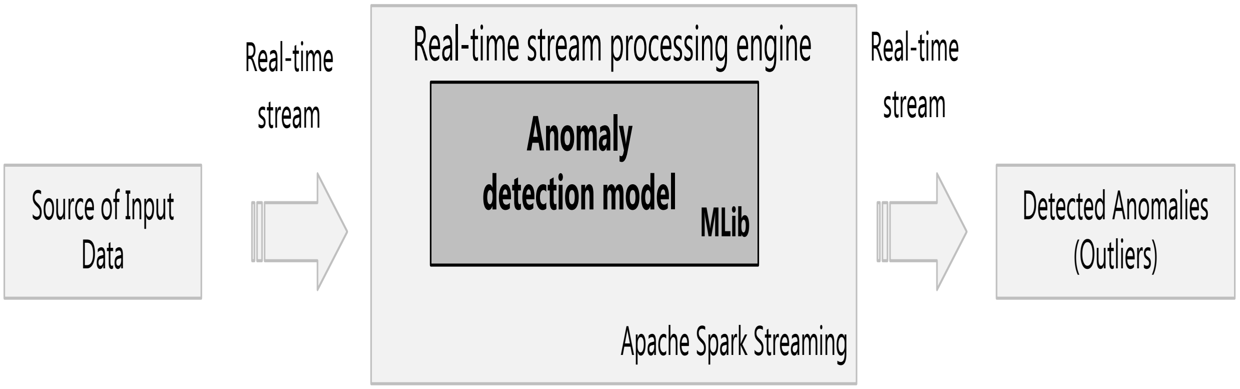

Figure 2: Real Time Outlier Detection Data Flow

Table of Summary Prepared with Help from Survey References (Agrawal and Agrawal, 2015), (Chandola *et al.*, 2009] and (Kaur and Singh 2016)

Techniques		**Nearest Neighbor Based Detection Techniques**	**Clustering-Based Anomalies Detection Techniques**	**Statistical Techniques**	**Classification Techniques**
Assumption	Normal data instances	present in dense neighbourhoods	belong to a cluster in the data, lie close to their closest cluster centroid, belong to large and dense clusters,	occur in high probability regions of a stochastic model	A classifier that can distinguish between normal and anomalous classes can be learnt in the given feature space.
	Anomalies	occur far from their closest neighbours	does not belong to any cluster, are far away from their closest cluster centroid, are either too small or too sparse clusters.	occur in the low probability regions of the stochastic model	
Advantages		• Unsupervised/semi-supervised mode • Simplest approach	• Unsupervised • Fast comparison • Can be adapted to complex data types	• Unsupervised • Confidence interval is provided with anomaly score • Histogram based techniques are simple	• Fast testing phase process • Improved efficiency with ensemble methods
Disadvantages		• High computational cost in testing phase • Difficult where several regions are with widely differing densities. • Difficult to identify in case if anomalies are present in groups. • Dependent on the proximity measures used	• High computation cost in cluster formation phase • A data object not belonging to any cluster may be a noise rather than an anomaly • Not suited for large datasets • Fail to label anomalies in certain cases	• Fail to label the anomalies correctly in certain cases • Difficult to find best statistic • For multivariate data it fails to capture the interactions between different attributes.	• Heavy dependency and reliability on training data • Class imbalance problem

Machine Learning Techniques Summary after Analyzing (Kaur and Singh, 2016; Sharma *et al.*, 2016; Omar *et al.*, 2013; Dunham, 2013; Hebbar and Mohan, 2015; Singh and Nene, 2013; Kaur *et al.*, 2013; Haq *et al.*, 2015; Asgharzadeh and Jamali, 2015; Agrawal and Agrawal, 2015; Shah *et al.*, 2015) Papers and Articles:

	Decision Tree	**SVM**	**Naive Bayes**	**ANN**	**Fuzzy Logic**	**GA**	**K-Means**
Technique	Classification	Classification & Regression	Classification	Classification	Classification	Classification	Clustering
Computation cost	High	High	Less	-	High	-	-
High dimensional data	-	Yes	Yes	Yes	-	-	-
Advantages	• Easy to understand for smaller trees • Requires little data preparation • Numerical &categorical data • Validate a model using statistical tests • High detection accuracy.	• Learning ability for small set of samples. • High training rate and decision rate, insensitiveness to dimension of input data	• Easy construction • Takes short computation time; • Works efficiently with large dataset	• Ability to generalize from limited, noisy and incomplete data. • Ease of use • Detect unknown intrusions. • Supports multiclass detection.	• Permits a data point to be in more than one cluster. It has a more natural representation of the behavior of genes. It's effective, especially against port scans and probes.	• Derives best classification rules. • Selects optimal parameters.	• Simple to use.
Disadvantage	• Optimal decision tree is NP-complete. • Complex trees do not generalize the data well. • Complex to large trees	• Positive &negative examples req. • High dependency on selecting good kernel function. • Training phase requires longer time.	• Difficult to handle continuous features. • Highly dependent on prior knowledge.	• Training required • Needs to be emulated. • Longer training process. • Over-fitting issue	• Need to determine membership cutoff value • Clusters are sensitive to initial assignment of centroids	• Can't assure constant optimization response times. • Over-fitting issue	• Necessity of specifying k. • Sensitive to noise • Clusters are sensitive to initial assignment of centroids.

For real time outlier detection system, it's eminent to establish live stream capture and process it and output the result in streaming to outlier detection notification.

As a first step towards building the real-time anomaly detection system, model needs to be build that will be used to predict anomaly. Below are sections describes the same.

c) Use Case Diagram

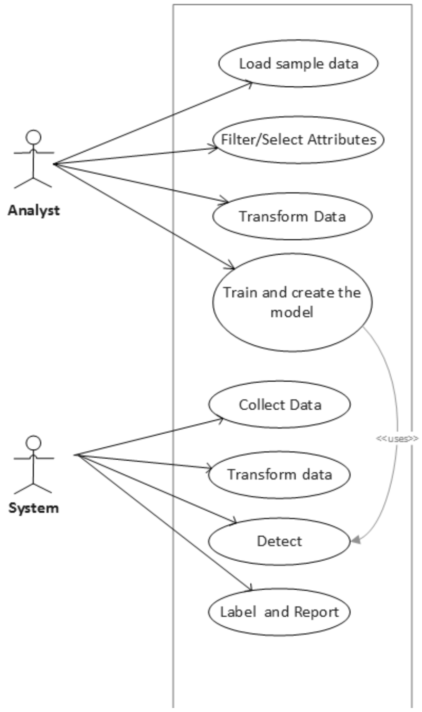

Figure 3: Use Case Diagram

d) Sequence Diagram

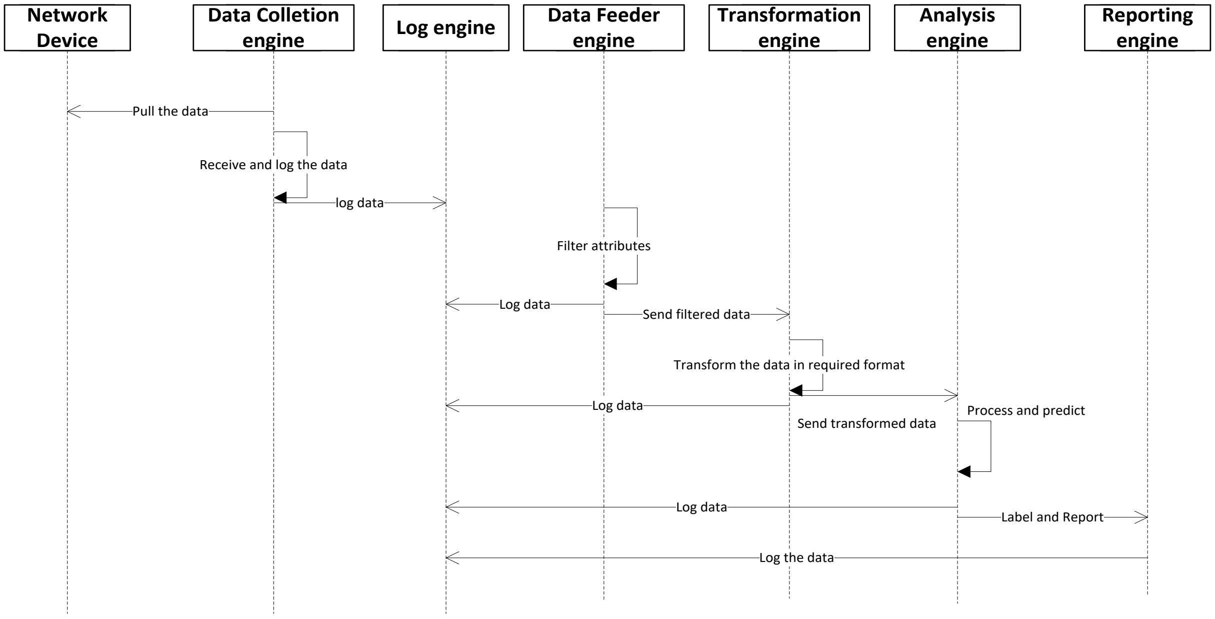

Figure 4: Sequence Diagram

e) Steps Planned to Perform for Model Generation and Evaluation

NSL KDD Dataset is planned to be used as it is publically available and relevant to our research area for model generation and evaluation.

Below are steps that needs to be followed:

1) Import data
2) Edit Metadata
3) Convert Indicator Values
4) Select Columns in dataset
5) Feature selection
6) Partition and sample
7) Use ‘Two-Class Logistic Regression” and “Boosted Decision Tree” on separate partitions
8) Tune parameters using metric for measuring performance for classification/regression
9) Score Model by adding scored labels and scored possibilities
10) Evaluate model using either of ROC(True Positive Rate Vs False Positive Rate)/Regression Vs Recall/LIFT (Number of True Positive Vs Positive Rate)
11) Compare performance and conclude which model to be used

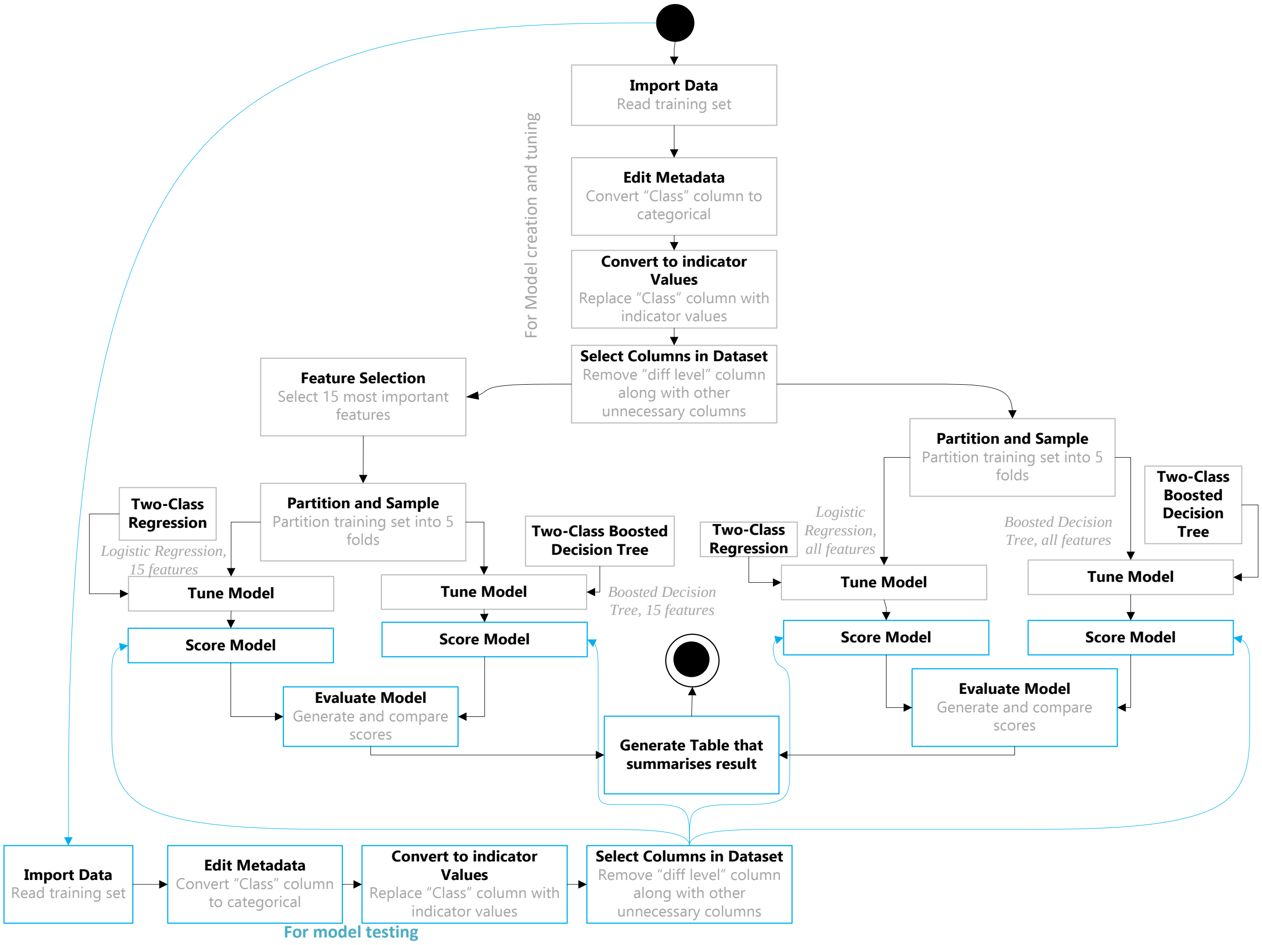
For Model creation and tuning
Import Data
Read training set
Edit Metadata
Convert “Class” column to categorical
Convert to indicator Values
Replace “Class” column with indicator values
Select Columns in Dataset
Remove “diff level” column along with other unnecessary columns
Feature Selection
Select 15 most important features
Partition and Sample
Partition training set into 5 folds
Two-Class Regression
Logistic Regression, 15 features
Tune Model
Score Model
Two-Class Boosted Decision Tree
Boosted Decision Tree, 15 features
Tune Model
Score Model
Evaluate Model
Generate and compare scores
Partition and Sample
Partition training set into 5 folds
Two-Class Regression
Logistic Regression, all features
Tune Model
Score Model
Two-Class Boosted Decision Tree
Boosted Decision Tree, all features
Tune Model
Score Model
Evaluate Model
Generate and compare scores
Generate Table that summarises result
Import Data
Read training set
Edit Metadata
Convert “Class” column to categorical
Convert to indicator Values
Replace “Class” column with indicator values
Select Columns in Dataset
Remove “diff level” column along with other unnecessary columns
For model testing

Figure 5: Activity Diagram

f) Boosted Decision Tree Classifier

The decision forest algorithm is an ensemble learning method for classification. Ensemble models provide efficient coverage and accuracy as compared to a single decision tree. The algorithm builds multiple decision trees and then voting is done on the most popular output class. Voting is used as a form of aggregation, wherein each tree in a classification decision forest outputs labels.

The aggregation process sums these results and normalizes them to get the "probability score" for each label. The trees having high prediction confidence/score will have a greater weight in the final conclusion of the ensemble from these multiple trees.

Decision trees are one of examples of non-parametric models, and they support data with varied distributions. Hence, they do not depend on any knowledge of distribution characteristics. For each tree,

Algorithm: Decision tree

```
Split(node, {example}):
1. A← the best attribute for splitting the {examples}
2. Decision attribute for this node ← A
3. For each value of A, create new child node
4. Split training {examples} to child nodes
5. For each child node/subset:
If subset is pure: STOP
Else: Split(node,{subset})
```

Calculate entropy for each attribute and decide which all attribute should be taken for building the model. In our model creation, attributes will be selected based on Chi Squared technique.

Following boosting method is used:

Algorithm: Boosted Decision Tree

```
1.    Start with an empty ensemble trees
2.    Sum of output from ensemble trees is taken for each of the training example
3.    Gradient of the loss function is calculated.
4.    Use gradient of loss function to fit a weak learner as the target function
5.    Add the weak learner to the ensemble and if required threshold not met, go to Step 2
```

g) Two-Class Logistic Regression

Two-Class Logistic Regression algorithm is used to create a model to predict one of two states.

It is a supervised learning method, and therefore its pre-requisite is to have labeled dataset with only two values-Normal or Anomalous.

Conclusion

As per the survey done of various papers pertaining to "Anomaly Detection with use of Machine Learning" and "Anomaly detection for DoS attack detection", it is learnt that:

- Algorithms and techniques based on distance measure and similarity measure are available for detecting anomalies;
- Specifically for DoS detection, data collection techniques and data analysis techniques have been in research area and seems to be suited for commercial use;
- Selection of algorithm depends on availability of labelled data, type of data, characteristics of data and the domain;
- So, to build effective Anomaly detection for DoS attack using machine learning, one has to have understanding of classification, clustering and statistical algorithms;
- For better rate (to have lower score) of false positive, ensemble technique can be used like first using principal component analysis (PCA) to reduce number of attributes/features and then using classification or clustering technique;
- There is no single rule book for effective anomaly detection, the input data attribute selection, building model and agreeing on the threshold impacts effectiveness of anomaly detection; &
- There are large numbers of open source tools available that may form part of component for building DoS IDS System using Machine Learning.

Next steps would be to suggest and setup platform for research in anomaly detection using open source tools for identifying anomaly using a sample dataset and further extend the same for real time streaming anomaly detection.

Few Open Source Tools are Listed Below that could be Probable Candidates for Platform Component in Proposed Model

Open Source Tool Name	Short Description
ELKI	K-means, anomaly detection, spatial index structures, apriori algorithm, dynamic time warping, and principal component analysis (PCA).
Rapid Miner	Machine learning, data mining, text mining, predictive analytics, and business analytics.
WEKA	Bayesian logistic, naive Bayes for text classification, Functional trees, Decision table/naive Bayes hybrid classifier, Rotation Forest, Multi-class alternating decision trees, K-means, Cobweb, DBScan, EM
Mahout	Premade algorithms for Scala + Apache Spark
WSO2	Decision Tree, K-Means, Regression, Support Vector Machine, Random Forest algorithm
Massive Online Analysis (MO A)	classification, regression, clustering, outlier detection, concept drift detection and recommender systems with tools for evaluation.
MEKA project	MEKA is based on the WEKA Machine Learning Toolkit.
Mallet	Naive bayes and decision tree
Encog	SVM,ANN, Genetic Programming, Bayesian Networks, Hidden Markov Models, Genetic Programming and GA.
Datumbox	Rapid development Machine Learning and Statistical applications.
Deeplearning4 j	Distributed deep-learning.
SNORT	Real-time traffic analysis and packet logging on IP networks

REFERENCES

Agrawal S and Agrawal J (2015). Survey on Anomaly Detection using Data Mining Techniques. In *International Conference on Knowledge Based and Intelligent Information and Engineering Systems*, **60** 708-713 Department of Computer Science and Engineering, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India.

Asgharzadeh P and Jamali S (2015). A survey on intrusion detection system based support vector machine algorithm. *International Journal of Research in Computer Applications and Robotics* **3**(12) 42-50.

Chandola V, Kumar V and Banerjee A (2009). Anomaly Detection : A Survey. *ACM Computing Surveys* **41**(3).

Gutierrez SA and Branch JW (2012). Application of Machine Learning Techniques to Distributed Denial of Service (DDoS) Attack Detection: A Systematic Literature Review.

Haq NF, Rafni M, Onik AR, Shah FM, Farid D and Hridoy AK (2015). Application of Machine Learning Approaches in Intrusion Detection System: A Survey. (IJARAI) *International Journal of Advanced Research in Artificial Intelligence* **4**(3) 9-19.

Hebbar R and Mohan K (2015). Network Attack Detection Using Machine Learning Approach. In *International Conference*, *"Computational Systems for Health & Sustainability", Bangalore,.*

Karami M and Kheyri D (2012). A Comprehensive Survey on Anomaly-Based Intrusion Detection. *Computer and Information Science* **5**(4) 132-140.

Kaur H, Minhas J and Singh G (2013). A review of Machine Learning based Anamoly Detection Techniques. *International Journal of Computer Applications Technology and Research* **2**(2) 185-187.

Kaur R and Singh S (2016). A survey of data mining and social network analysis based anomaly detection Techniques. *Egyptian Informatics Journal* **17** 199–216.

MH Dunham (2013). Data Minig, Pearson, 2013.

N. P. f. Intelligent, "Numenta," (2015) [Online]. Available: https://numenta.com/assets/pdf/whitepapers/Numenta%20White%20Paper%20-%20Science%20of%20Anomaly%20Detection.pdf.

Omar S, Ngadi A and Jebur HH (2013). Machine Learning Techniques for Anomaly Detection: An Overview. *International Journal of Computer Applications (0975 8887)* **79**(2).

Rao KH (2011). Implementation of Anomaly Detection Technique Using Machine Learning Algorithms. *International Journal of Computer Science and Telecommunication* **2**(3) 25-31.

Shah AA, Khiyal MSH and Awan MD (2015). Analysis of Machine Learning Techniques for Intrusion Detection System: A Review. *International Journal of Computer Applications* **119**(3) 19-40.

Sharma N, Mahajan A and Mansotra V (2016). Machine Learning Techniques Used in Detection of DOS Attacks: A Literature Review. *International Journal of Advanced Research in Computer Science and Software Engineering* **6**(3) 100-106.

Singh J and Nene MJ (2013). A Survey on Machine Learning Techniques for Intrusion Detection Systems. *International Journal of Advanced Research in Computer and Communication Engineering* **2**(11).

Weller-Fahy DJ, Borghetti BJ and Sodemann AA (2015). A Survey of Distance and Similarity Measures Used Within Network Intrusion Anomaly Detection. *IEEE Communication Surveys & Tutorials* **17**(1) 22.