# Analyzing the Behavior of Multiple Dimensionality Reduction Algorithms to Obtain Better Accuracy using Benchmark KDD CUP Dataset

Suriya Prakash Jambunathan[1], Suguna Ramadass[2], and Palanivel Rajan Selva kumaran[3]
[1]Faculty of Information and Communication Engineering, Anna University, India
[2]Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, India
[3]Department of Electronics and Communication Engineering, M. Kumarasamy College of Engineering, India

**Abstract:** *In the ubiquitously connected world of IT infrastructure, Intrusion Detection System (IDS) plays vital role. IDS is considered as a critical component of security infrastructure and is implemented either through hardware or software devices and can detect malicious activities in a networked environment. To detect or prevent network attacks, Network Intrusion Detection (NID) system may be equipped with machine learning algorithms to achieve better accuracy and faster detection speed. Analyzing different attacks effectively through Dimensionality Reduction Algorithms is an efficient mechanism. The significance of these algorithms is they improvise feature selection from huge datasets. Also through this the learning speed is enhanced. Speed is a crucial parameter in the success of network intrusion detection systems for defending reactions. In this paper open source datasets Knowledge Discovery in Databases (KDD CUP) dataset and 10% KDD CUP dataset are employed for experimentation. These datasets are provided to Dimensionality Reduction Algorithms like Principal Component Analysis (PCA), Linear Discriminate Analysis (LDA) and Kernel PCA with different kernels and classified with Logistic Regression classification algorithm for procuring accurate results. Further to boost up the accuracy achieved so far K-fold algorithm is utilized. Finally a comparative study of different accuracy results is done by using K-fold algorithm and also without the usage of this algorithm. The empirical study on KDD CUP data confirms the effectiveness of the proposed scheme. In this paper we discovered the combination of multiple dimensionality reduction algorithm such as PCA , LDA and Kernel PCA with classification algorithm and this combination of algorithm gives best result. Our study will help out the researchers to uncover critical area such as intrusion detection in network traffic environment. The results what we identified will be very much helpful for researchers for their future research on KDD CUP dataset. In this the new theory will be arrived by this research that the best accuracy achieved by PCA with 10% KDD CUP dataset experimental results without KFold attained 98% and with KFold attained 99%. LDA with 10% KDD CUP Dataset experimental results without KFold attained 98% and with KFold attained 99%.*

**Keywords:** *Intrusion attacks, network, features, accuracy.*

*Received December 14, 2020; accepted August 17, 2021*
*https://doi.org/10.34028/iajit/19/1/14*

## 1. Introduction

The obtained accuracy is predicted using detection rate of intrusions. The feature selection algorithm is employed to predict the exact features and it also supports by minimizing the computation time. The KDD CUP 1999 intrusion dataset is an internationally accepted benchmark intrusion dataset. In data mining and machine learning, feature selection plays a major role in fundamental statistical techniques. In different verticals like finance, biology, social science, plethora of irrelevant features creep in [1]. A novel Deep-Feature Extraction And Selection (D-FES) combining stacked feature extraction and feature selection is proposed. By extracting the raw input and meaningful, relevant features through stacked auto encoding feature selection model the authors are minimizing the load on machine learning model and computational complexity is also minimized [2]. This leads to dimensionality curse and demands for reduction of dimensionality in feature selection process. Thus, the challenge of identifying most characterizing features and eliminating irrelevant and redundant features is to be addressed. During feature selection process determining feature sorting techniques has significance. Also correlation measures have been widely used to achieve feature selection [3]. Feature subset model envisaged in form of probabilistic representations emphasizes the need for repeated refines, model extraction and evaluation. By this process the dataset is split vertically based on the distributed optimization method and thus unbalanced dimensions are eliminated [4]. To compute MC

efficiently, Alternating Conditional Expectation (ACE) is proposed by Breiman and Friedman. However, the univariate feature selection results in mismatch in most feature selection problems thus authors were motivated to consider connected relationship between variables [5]. The Intrusion Detection System (IDS) analyzes the available security protection methods for computers and the networks. In this system, two types of attack methods such as misuse and anomaly are used. The misuse approach operates on known attacks and identifies the intrusions by matching them with the existing patterns or signature. Anomaly detection system identifies the normal behavior by monitoring the system and raises an alarm and reports on occurrence of deviations in the network behavior [6, 13]. Authors proposed algorithm to detect the intrusion and join with the statistical approach model, to produce efficient result [7]. Winding faults in induction motors is addressed using multiple feature extraction techniques also most relevant features were extracted through this. Next authors employed neural network classifiers to classify the extracted features [8].

### 1.1. Major Contribution

1. This research work proposes to identify better accuracy by analyzing different Dimensionality Reduction Algorithm combined with classification algorithm.
2. Advanced dimensionality reduction techniques have been applied to get better accuracy results.
3. A novel intrusion detection accuracy analysis framework has been proposed to set the benchmark for the researcher.
4. A novel intrusion detection algorithm has been proposed to illustrate the process of suspicious flow detection analysis.
5. Experiments have been conducted on the benchmark KDDCUP and 10% KDDCUP datasets. The results show that the proposed model outperformed on both datasets.
6. We compared the combination of different dimensionality reduction algorithm and classification algorithm combinations, which shows the greater performance than other models.

The rest of this paper is arranged as follows. Section 2 presents a literature survey of previous works and models related to the intrusion detection in network traffic. Section 3 describes the methodology of the research work, which includes one novel framework and novel algorithm of intrusion detection in network traffic approach. Section 4 discusses the experimental results of the proposed model. Finally, section 5 concludes the research and outlines future work.

## 2. Literature Survey

In the real world, multiple applications run on the computer networks and network security cannot be compromised. Intrusion detection plays vital role in security of network infrastructure and it is in demand for managing the network security. Security goals are determined by three factors, i.e., availability, integrity and confidentiality [9]. Authors proposed two models: First, parallelized the Direct Matching Algorithm (PDMA) and Second, the PDMA implemented in Network Intrusion Detection Systems (NIDS) to enhance the speed of the NIDS detection engine [10]. A work on relaxing the information content for computational efficiency by an unsupervised feature selection method is proposed and the best feature subset by minimal computational cost is achieved [11]. Research on multiple problems in feature selection proposes usage of new software named Easy-to-use and Standalone Software (ECoFFeS) and also evolutionary algorithms are brought into play. ECoFFeS software is an Application Programming Interface that interrelates single object, multi-object evolutionary algorithms and also regression and classification models [12].

Table 1. Comparative analysis with existing methods.

| Authors and Year | Methodology | Sensitivity (%) | Specificity (%) | Accuracy (%) |
|---|---|---|---|---|
| Proposed Methodology (In this paper) | DR+LR classification approach | 98.1 | 93.4 | 96.2 |
| Hnaif *et al.* [10] | Parallel Scalable Approximate Matching Algorithm | 94.2 | 90.8 | 92.6 |
| Brankovic *et al.* [4] | A distributed feature selection algorithm | 92.1 | 84.7 | 90.6 |
| Aminanto *et al.* [2] | Deep Abstraction and Weighted Feature Selection | 90.5 | 91.7 | 91.6 |
| Biesiada and Duch [3] | Feature selection for high dimensional data | 90.7 | 82.1 | 87.5 |

Stratified Feature Ranking (SFR) is employed for the high dimensional data by Subspace Feature Clustering (SFC). It identifies the importance of every feature in each class using feature clusters. These feature clusters are separated and ranked based on subspace weight [14]. Toloueiashtian *et al.* [15] proposed algorithm tries to find the Best Solution (BS) based on three operations and further compared with Genetic Algorithm (GA) and Ant Colony Optimization (ACO) based on time complexity criteria. KDD CUP 1999 dataset is a threshold dataset in identifying suspicious detection. It has 3 categorical and 38 numerical attributes totaling to 41 attributes and is labeled with special kind of attacks. These attacks can

be categorized into four types namely Denial of Service (DoS), User to Root (U2R) attack, Remote to local (R2L) and Probe attack.

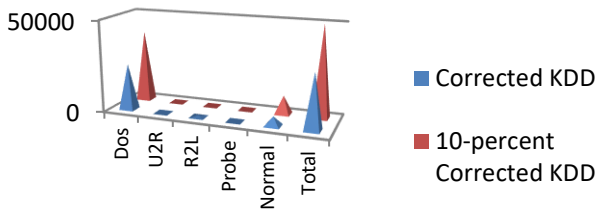


Figure 1. Distribution of KDD dataset.

DoS interrupts the access of the network, i.e., the authorized user will not be able to access the network and in U2R attack the intruder will access the root user access of the network. Using R2L attack the intruder will find out the vulnerable points in the system and try to access the system. In Probe attack, the intruder sends a blank message to identify whether the destination exists or not. Figure 1 describes the attack distribution of Corrected KDD dataset and 10 per cent KDD cup data sets.

## 3. Methodology

Figure 2 describes Un-supervised and Semi-Supervised architecture. Here KDD CUP and 10 % of KDD CUP datasets are imported. Subsequently the categorical dataset is extracted by removing the irrelevant features and by encoding the dependent variables. Next the dataset is split into Training set and Testing Set. Equations (1) and (2) denotes dimensionality reduction on Principal Component Analysis (PCA)

$$\frac{1}{n}\sum_{i=1}^{n}(\overrightarrow{w}.\overrightarrow{x_i})^2 = (\frac{1}{n}x^i - \overrightarrow{w})^2 + \mathrm{Var}[\overrightarrow{w}.\overrightarrow{x_i}] \quad (1)$$

$$\sum_{j=1}^{k}(\vec{x}_i - \overrightarrow{w_j})^2 \quad (2)$$

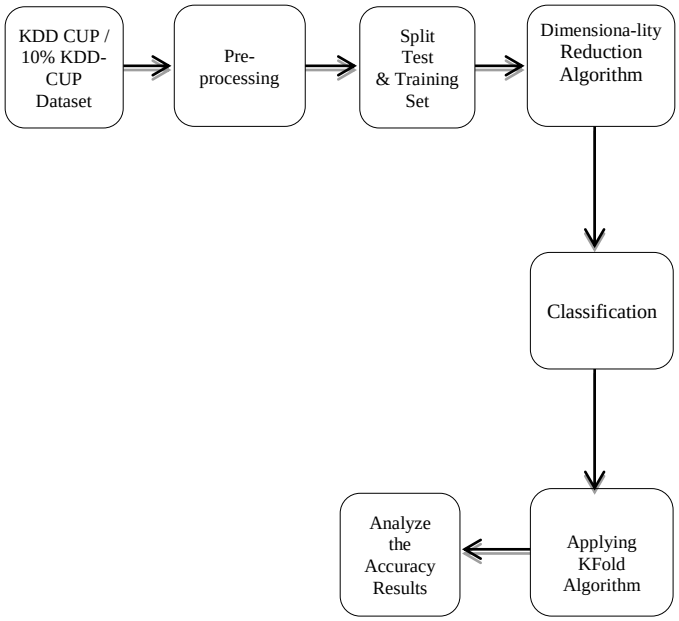


Figure 2. System architecture.

Further feature scaling in the Training set and the Testing set are accomplished to streamline the data with uniform value range. Followed by application of PCA or Linear Discriminate Analysis (LDA) for Unsupervised classification algorithm and Kernel PCA for Semi-Supervised and Logistic Regression classification algorithm used to classify the data and generate the confusion matrix. Equations (3) and (4) denotes Logistic Regression classification Algorithm.

$$p(X) = e^{(b_0 + b_1 * X)/(1 + e^{(b_0 + b_1 * X)})} \quad (3)$$

$$\log(\frac{p(x)}{1} - p(x)) = b_0 + b_1 + x \quad (4)$$

With these steps of processing data accuracy is predicted. Finally K-Fold algorithm is applied to improve the accuracy.

### 3.1. System Model

The existing feature selection algorithms do not guarantee the elimination of redundant variables. As the features are enormously huge in number in KDD cup dataset this paper proposes a new technique to achieve accurate results by avoiding the redundant variables in the dataset. The proposed method deals with KDD CUP and 10% KDD CUP dataset. The KDD cup dataset contains 12 classes and 19 attributes. The classes are named as back, buffer-overflow, ftp_write, guess_passwd, imap, ipsweep, land, load module, multi hop, neptune, namp and normal and there are 19 attributes that are said to be duration, dst_bytes, urgent, num_failed _logins, num_compromised, su_attempted, num_file_ creations, num_access_files, is_host_login, count ,serror_rate,reror_rate,same_srv_rate,srv_diff_host_rate,dst_host_srv_count,dst_host_diff_srv_rate,dst_host_srv_diff_host_rate,dst_host_srv_serror_rate, dst_host_srv_ rerror_rate.

These attributes and classes are identified by inspecting the packets rigorously for each flow. We imported the KDD CUP and 10 % KDD CUP dataset to the system model and then obtained Categorical Data by eliminating irrelevant features in the dataset. In next step the data is split into the training set and testing set and the Training set is Fa={Fa1, Fa2,. . .} with class La={La1, La2, . . .} and Testing flow set named as Ea={Ea1, Ea2, …} . Next step feature scaling is achieved by bringing uniformity in the data flow of every column of Training and Testing data. We applied PCA, LDA, and Kernel PCA algorithms which fall under Dimensionality Reduction algorithms category. Logistic Regression algorithm used in our experiment is a classification algorithm. Initially PCA is combined with Logistic Regression followed by LDA with Logistic Regression and Kernel PCA with Logistic Regression respectively. Finally to boost up the accuracy rate we employed K-Fold algorithm. Accuracy results are computed by using K-Fold

algorithm based analysis and by excluding K-Fold algorithm in the analysis.

*Algorithm 1: Minimal Feature selection Algorithm*

*Input: Minimal Training flow set F with its class L and Individual column Fc; Fcv represents each individual Value in each column of Fc; Maximal Test flow set E and its individual row Er; Erv represents each Individual value in Er.*
*Output: Correctly Classified set z and its class zl*
*1.Import the KDD CUP and 10% KDD CUP dataset D*
*2. C € D // Finding Categorical data in D*
*// Step 3 Splitting Training and Test Flow*
*3. L ← F; // Training set*
*4. Le← L; // Class*
*5.Applying PCA or LDA or Kernel PCA Algorithm*
*6.Applying Logistic Regression Classification Algorithm*
*7.Accuracy Boost up using Kfold Algorithm*
*8. Identify the Confusion Matrix to calculate the Accuracy*
*9. End*

- Stepwise Description of the Algorithm
  - *Step* 1: The algorithm considers KDD CUP dataset and 10% KDD CUP dataset denoted by D separately to obtain following features.
  - *Step* 2: Initially Categorical Data C is extracted from D.
  - *Step* 3: The extracted C is bifurcated into Training and Test flow dataset.
  - *Step* 4: The Training Set data L is further filtered to obtain class Le.
  - *Steps* 5 and 6: The Dimensionality Reduction Algorithm PCA and Logistic Regression Classification algorithm are applied on the data.
  - *Step* 7: Further Kfold algorithm is applied along with PCA and Logistic Regression algorithm to obtain better accuracy.
  - *Step* 5: and *Step* 6 and *Step* 7: These steps are repeated for LDA and Kernel PCA algorithms too.
  - *Step* 8: Confusion matrix is identified to calculate the accuracy.
  - *Step* 9: End of the algorithm.

## 4. Results and Discussion

The Feature selection algorithms employed in this work are Logistic Discriminant Analysis (LDA),

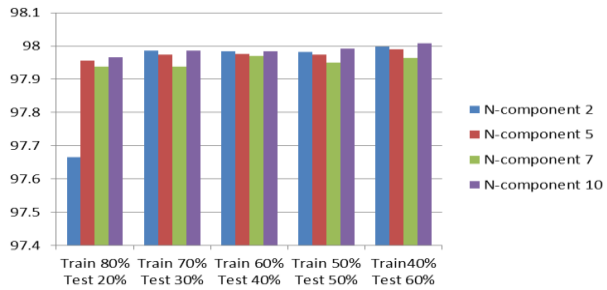


Figure 3. KDD CUP dataset using PCA.

Principal Component Analysis (PCA) and Kernel PCA. The paper identifies minimal features used to reduce the computation time and the results are depicted in the following graphs. Also it represents the optimal feature selection for KDD CUP dataset and 10 percent KDD CUP through feature selection algorithms achieved with good accuracy rate through comparative study. PCA dimensionality reduction algorithm is applied on comprehensive KDD CUP dataset. In the experiment the training and testing set are taken in the ratio of 80:20, 70:30,60:40, 50:50 and 40:60. Also N-components with varying component values are used. Figure 3 depicts the same. The inference drawn is 40:60 ratio of training versus testing dataset with N-component value-10 yields accuracy above 98%.PCA dimensionality reduction algorithm is used in combination of K-Fold algorithm to improve the accuracy on KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 4 depicts the same. The inference drawn is irrespective of the ratio of training versus testing datasets the N-component value-10 yields accuracy above 99%.

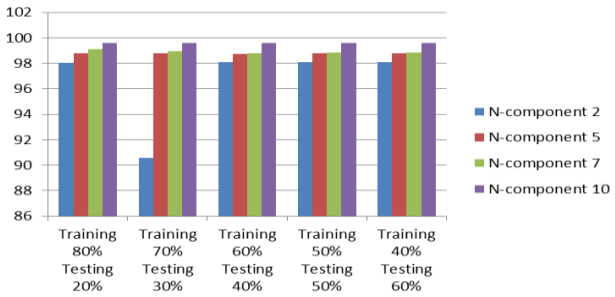


Figure 4. KDD CUP dataset using PCA with K-Fold.

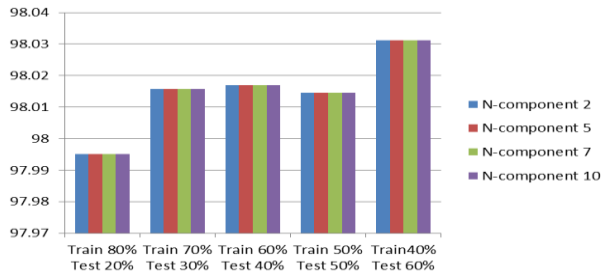


Figure 5. KDD CUP dataset using LDA.

LDA dimensionality reduction algorithm is applied on comprehensive KDD CUP dataset.

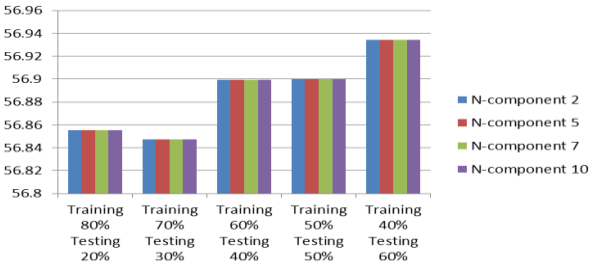


Figure 6. KDD CUP dataset using LDA with K-Fold.

In the experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 5 depicts the same. The inference drawn is for 40:60 ratio of training versus testing dataset irrespective of N-component values the accuracy obtained is 98.03%. LDA dimensionality reduction algorithm is used in combination of K-Fold algorithm to improve the accuracy on KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 6 depicts the same. The inference drawn is for 40:60 ratio of training versus testing datasets irrespective of the N-component values accuracy of 56.93% is obtained.

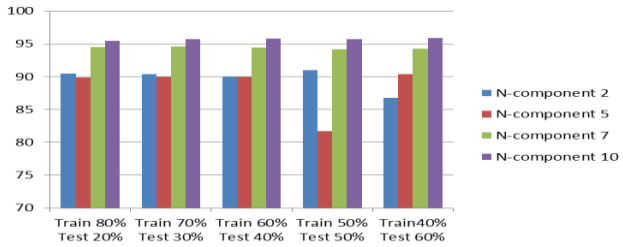


Figure 7. KDD CUP dataset using kernel PCA with RBF kernel.

Kernel PCA dimensionality reduction algorithm having Radial Basis Function (RBF) kernel is used on KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 7 depicts the same as explained above.

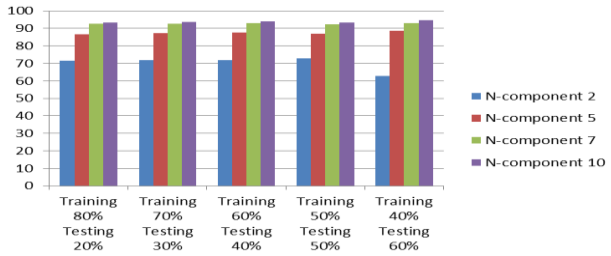


Figure 8. KDD CUP dataset using kernel PCA and using RBF kernel with KFold.

The inference drawn is 96% accuracy is obtained for all the ratios of training versus testing datasets with N-component value as 10.

Kernel PCA dimensionality reduction algorithm having RBF kernel with K-Fold algorithm for improving the accuracy is used on KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60 respectively. Also N-components with varying component values are used. Figure 8 depicts the same. The inference drawn is 95% accuracy is obtained for 40:60 ratios of training versus testing datasets with N-component value as 10.

Kernel PCA dimensionality reduction algorithm having Linear Kernel is used on KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 9 depicts the same. The inference drawn is 96% accuracy is obtained for 40:60 ratio of training versus testing datasets with N-component value as 5.

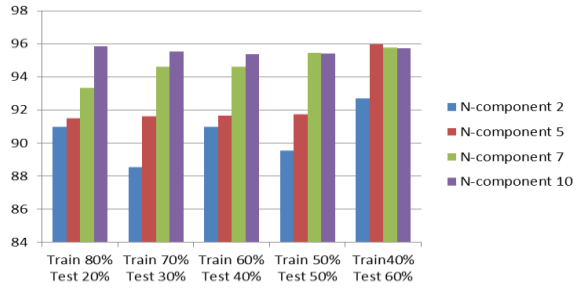


Figure 9. KDD CUP dataset using kernel PCA and using logistic kernel.

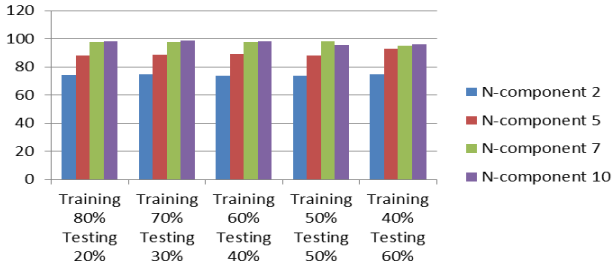


Figure 10. KDD CUP dataset using kernel PCA and using linear kernel with K-Fold.

Kernel PCA dimensionality reduction algorithm having Linear Kernel with K-Fold algorithm is used to improve the accuracy on KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 10 depicts the same. The inference drawn is 98% accuracy is obtained for 80:20 ratios of training versus testing datasets with N-component value as 10.

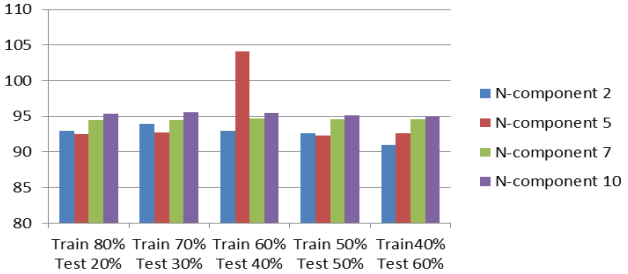


Figure 11. KDD CUP dataset using kernel PCA and using sigmod kernel.

Kernel PCA dimensionality reduction algorithm having Sigmod kernel is used on KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 11 depicts the same. The inference drawn is accuracy exceeds 100% for 60:40 ratios of training versus testing datasets with N-component value as 5.

Kernel PCA dimensionality reduction algorithm having Sigmod kernel with K-Fold to improve accuracy is used on KDD CUP dataset.

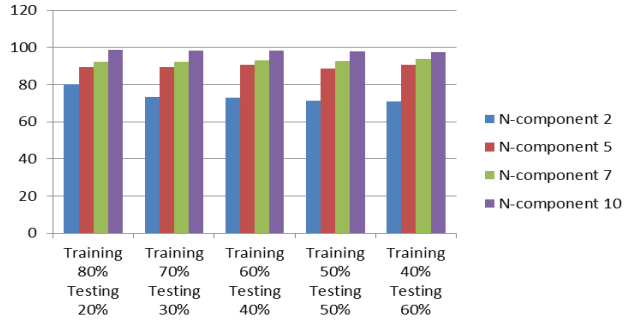


Figure 12. KDD CUP dataset using kernel PCA and using sigmod kernel using K-Fold.

In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 12 depicts the same. The inference drawn is 98% accuracy is obtained for all the ratios of training versus testing datasets with N-component value as 10.

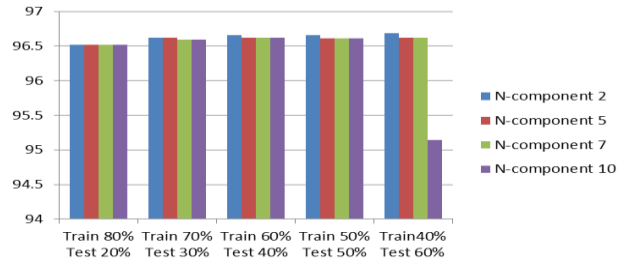


Figure 13. KDD CUP dataset using kernel PCA and using poly kernel.

Kernel PCA dimensionality reduction algorithm having Poly kernel is used on KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 13 depicts the same. The inference drawn is 96.8% accuracy is obtained for the 40:60 ratios of training versus testing datasets with N-component value as 2.

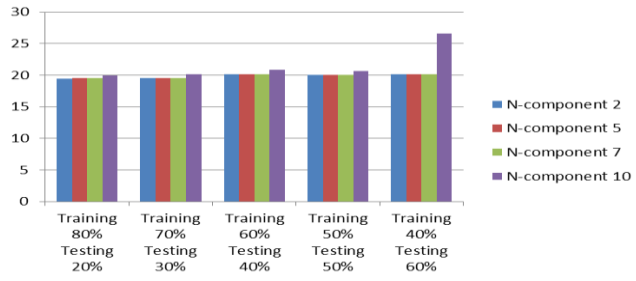


Figure 14. KDD CUP dataset using kernel PCA and using poly kernel with K-Fold.

Kernel PCA dimensionality reduction algorithm having Poly kernel with K-Fold is used to improve accuracy on KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 14 depicts the same. The inference drawn is 26% accuracy is obtained for the 40:60 ratios of training versus testing datasets with N-component value as 10.

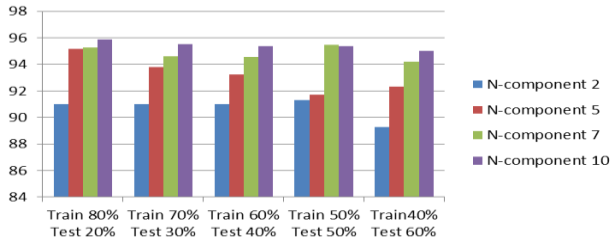


Figure 15. KDD CUP dataset using Kernel PCA and using cosine kernel.

Kernel PCA dimensionality reduction algorithm having Cosine kernel is applied on KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 15 depicts the same. The inference drawn is 95.8% accuracy is obtained for the 80:20 ratios of training versus testing datasets with N-component value as 10.

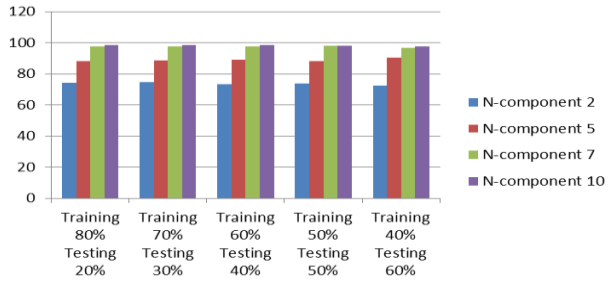


Figure 16. KDD CUP dataset using Kernel PCA and using cosine kernel with K-Fold.

Kernel PCA dimensionality reduction algorithm having Cosine kernel is combined with K-Fold algorithm for improved accuracy and is applied on KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 16 depicts the same. The inference drawn is 98% accuracy is obtained for all the ratios of training versus testing datasets with N-component value 10. PCA dimensionality reduction algorithm is applied on 10% KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 17 depicts the same.

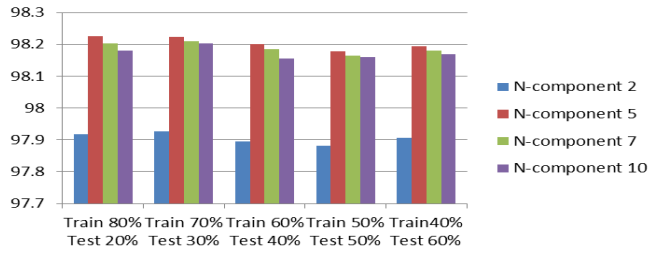


Figure 17. 10% KDD CUP dataset using PCA.

The inference drawn is 98.2% accuracy is obtained for 80:20 ratios of training versus testing datasets with N-component value 5.


Figure 18.10% KDD CUP dataset using PCA with KFold.

PCA dimensionality reduction algorithm is combined with K-Fold algorithm for improved accuracy and is applied on 10% KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 18 depicts the same. The inference drawn is 99.1% accuracy is obtained for 60:40 ratio of training versus testing datasets with N-component value 10.

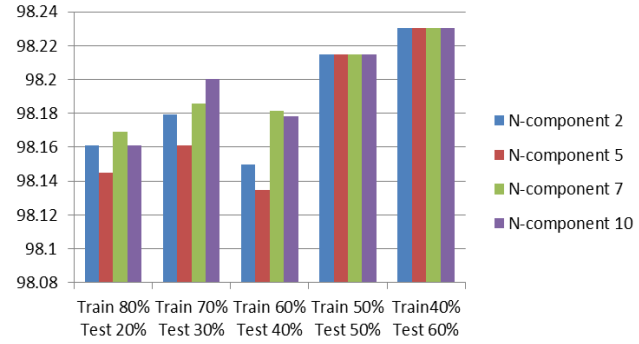


Figure 19. 10% KDD CUP dataset using LDA.

LDA dimensionality reduction algorithm is combined with K-Fold algorithm for improved accuracy and is applied on 10% KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 19 depicts the same. The inference drawn is 98.24% accuracy is obtained for 40:60 ratio of training versus testing datasets with N-component values 2, 5, 7, and 10 respectively.

LDA dimensionality reduction algorithm is combined with K-Fold algorithm for improved accuracy and is applied on 10% KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60.


Figure 20. 10% KDD CUP dataset using LDA with KFold.

Also N-components with varying component values are used. Figure 20 depicts the same. The inference drawn is 99% accuracy is obtained for 80:20,70:30, and 60:40 ratio of training versus testing datasets with N-component value 10.

Kernel PCA dimensionality reduction algorithm using cosine kernel is applied on KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 21 depicts the same.

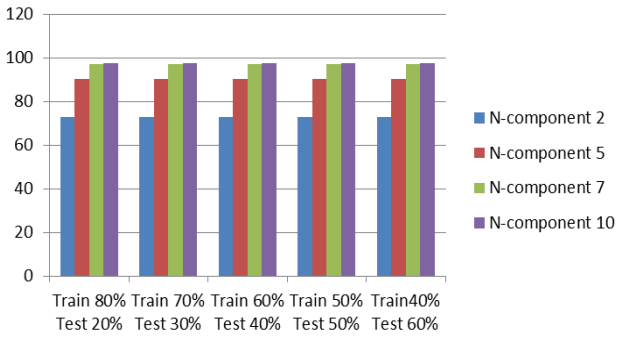


Figure 21. 10% KDD CUP dataset using Kernel PCA with cosine kernel.

The inference drawn is 99% accuracy is obtained for all the ratios of training versus testing datasets with N-component value 10.


Figure 22. 10% KDD CUP dataset using Kernel PCA with cosine kernel with KFold.

Kernel PCA dimensionality reduction algorithm using cosine kernel is combined with K-Fold algorithm for improved accuracy and is applied on KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 22 depicts the same. The inference drawn is 99% accuracy is obtained for all ratio of training versus testing datasets with N-component value 10.

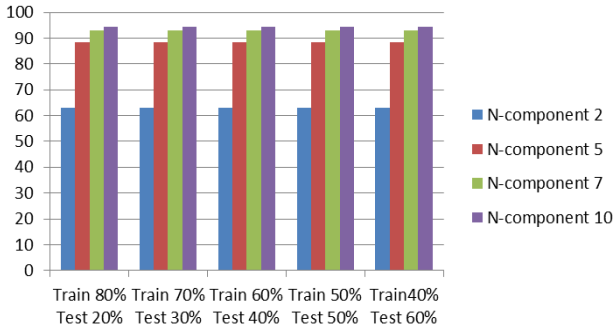


Figure 23. 10% KDD CUP dataset using kernel PCA with RBF kernel.

Kernel PCA dimensionality reduction algorithm using RBF Kernel is applied on 10% KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 23 depicts the same. The inference drawn is 95% accuracy is obtained for all ratio of training versus testing datasets with N-component value 10.

Kernel PCA dimensionality reduction algorithm with rbf kernel is combined with K-Fold algorithm for improved accuracy and is applied on 10% KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60.

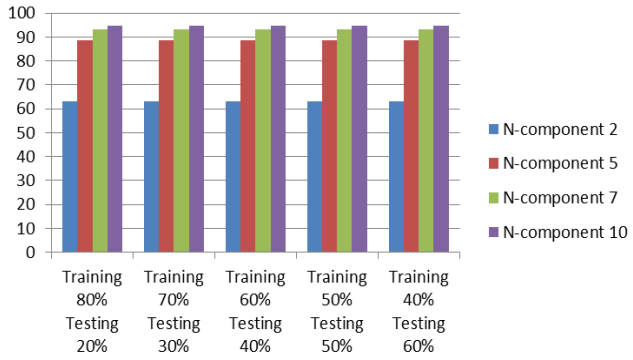


Figure 24. 10% KDD CUP dataset using Kernel PCA with RBF Kernel with KFold.

Also N-components with varying component values are used. Figure 24 depicts the same. The inference drawn is 95% accuracy is obtained for all ratios of training versus testing datasets with N-component value 10.

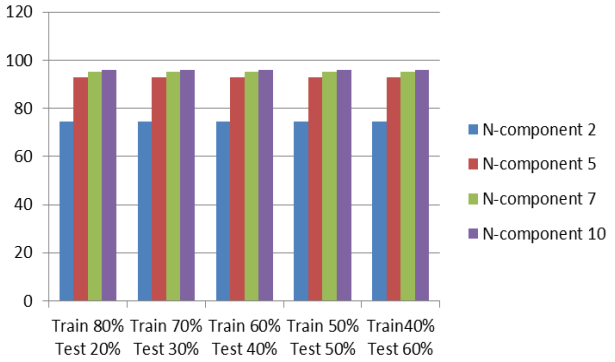


Figure 25. 10% KDD CUP dataset using kernel PCA with linear kernel.

Kernel PCA dimensionality reduction algorithm using Linear Kernel is applied on 10% KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 25 depicts the same. The inference drawn is 96% accuracy is obtained for 80:20 ratio of training versus testing datasets with N-component value 10.

Kernel PCA dimensionality reduction algorithm using Linear Kernel is combined with K-Fold algorithm for improved accuracy and is applied on 10% KDD CUP dataset.


Figure 26. 10% KDD CUP dataset using kernel PCA with linear kernel with KFold.

In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 26 depicts the same. The inference drawn is 96% accuracy is obtained for all ratio of training versus testing datasets with N-component value 10.

Kernel PCA dimensionality reduction algorithm using sigmod kernel and is applied on KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60.

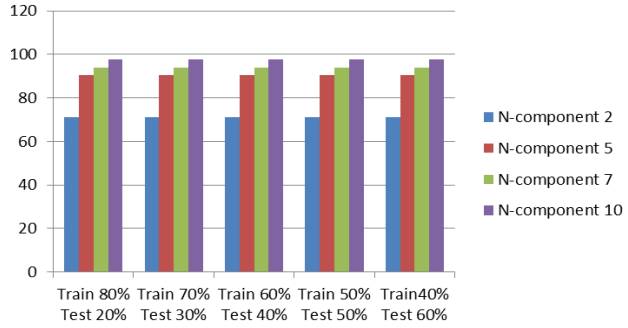


Figure 27. 10% KDD CUP dataset using kernel PCA with sigmod kernel.

Also N-components with varying component values are used. Figure 27 depicts the same. The inference drawn is 99.4% accuracy is obtained for all the ratio of training versus testing datasets with N-component value 10.

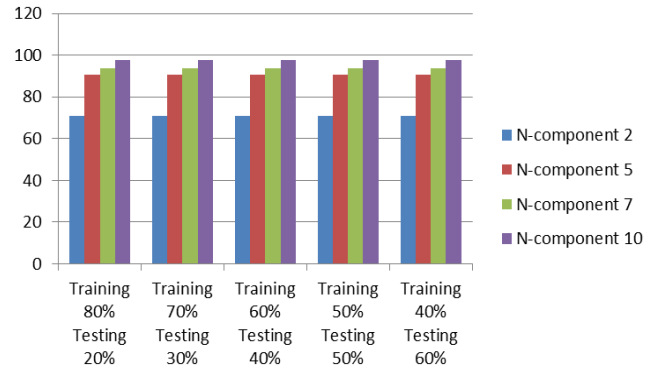


Figure 28. 10% KDD CUP dataset using Kernel PCA with sigmod Kernel with KFold.

Kernel PCA dimensionality reduction algorithm with sigmod kernel is combined with K-Fold algorithm for improved accuracy and is applied on 10% KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 28 depicts the same. The inference drawn is 99% accuracy is obtained for all ratios of training versus testing datasets with N-component value 10.

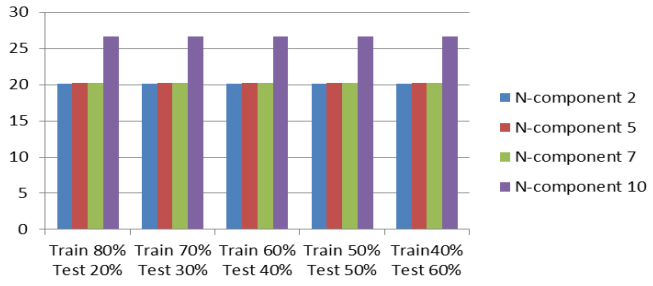


Figure 29. 10% KDD CUP dataset using Kernel PCA with poly kernel.

Kernel PCA dimensionality reduction algorithm with poly kernel is applied on 10% KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 29 depicts the same. The inference drawn is 27% accuracy is obtained for all the ratios of training versus testing datasets with N-component value 10.


Figure 30. 10% KDD CUP dataset using Kernel PCA with poly kernel with KFold.

Kernel PCA dimensionality reduction algorithm with poly kernel is combined with K-Fold algorithm for improved accuracy and is applied on 10% KDD CUP dataset. In this experiment the training and testing set are taken in the ratio of 80:20, 70:30, 60:40, 50:50, and 40:60. Also N-components with varying component values are used. Figure 30 depicts the same.

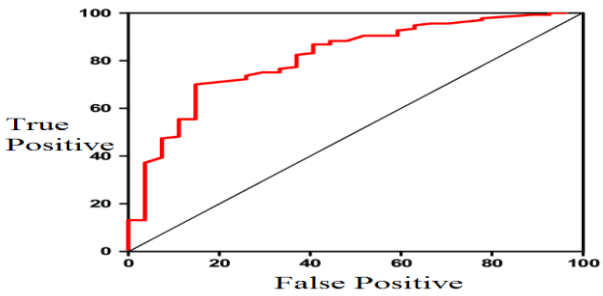


Figure 31. (ROC) curve is the plot that depicts the true positive and false positive.

The inference drawn is 26% accuracy is obtained for all ratios of training versus testing datasets with N-component value 10. In this work the computational cost respect to execution time per time step during simulation. The results show that precision increased by 4.56%, total computation decreased by 7.54%, and detection speed increased by 7%, which basically realized the requirement of real time high-precision detection.

Receiver Operating Characteristic (ROC) curve is the plot that depicts the true positive and false positive by different combination of dimensionality reduction algorithm with classification algorithms.

In this paper, We worked with open source software named as anaconda navigator IDE in that spyder scientific environment to complete this work. There is no cost involved in implementing the proposed approach because of open source software and we

identified in combination of dimensionality reduction algorithm and classification algorithm gives better accuracy in KDD CUP and 10% KDD CUP dataset. We identified that in KDD CUP dataset using PCA and Logistic regression combination 50% Training set and 50% Test set split-up achieved accuracy as 97% with kfold and 98% without kfold and using LDA and Logistic regression combination 40% Training set and 60% Test set split-up achieved accuracy as 97% with kfold. Kernel PCA with poly Kernel and Logistic regression combination 60% Training set and 40% Test set split-up achieved accuracy as 96% with kfold. Kernel PCA with linear Kernel and Logistic regression combination 80% Training set and 20% Test set split-up achieved accuracy as 95% with kfold and 98% without kfold. Kernel PCA with sigmod Kernel and Logistic regression combination 60% Training set and 40% Test set split-up achieved accuracy as 95% with kfold and 98% without kfold. Kernel PCA with RBF Kernel and Logistic regression combination 40% Training set and 60% Test set split-up achieved accuracy as 95% with kfold and 98% without kfold. Kernel PCA with cosine Kernel and Logistic regression combination 50% Training set and 50% Test set split-up achieved accuracy as 95% with kfold and 98% without kfold.

## 5. Conclusions

This paper proposes a novel methodology for classifying the intrusion using different classification methods. The entire work of this paper emphasizes on the minimum feature selection techniques applied to KDD CUP dataset and 10% KDD CUP dataset. A comparative analysis is done with three different dimensionality reduction algorithms and Logistic Regression classification algorithm with and without K-Fold algorithm to boost up the accuracy. The three Dimensionality Reduction Algorithms employed in the experiments are PCA, LDA and Kernel PCA. Moreover five different Kernels of Kernel PCA namely cosine, sigmod, linear, RBF and poly kernels are used in this research. The Confusion Matrix results are used to obtain the accuracy. The important conclusion drawn from our experiments and the proposed model is: The comparative results for comprehensive KDD CUP dataset and 10% KDD CUP dataset is obtained. In KDD CUP dataset the overall analysis of all the results are depicted and the conclusion we could arrive at is that the 95% accuracy is obtained by using the ratio of 40:60 with N-component value as 10. While in the 10% KDD CUP dataset the overall analysis of all results conclusion drawn is that the 99% accuracy is obtained by using a ratio of 40:60 with N-component value as 10. The future enhancement we are contemplating is to employ different classification algorithms to compute better accuracy results.

## Reference


[1] Albert R., "Network Inference, Analysis, and Modeling in Systems Biology," *The Plant Cell*, vol. 19, no. 11, pp-3327- 3338, 2007.

[2] Aminanto M., Choi R., Tanuwidjaja H., Yoo P D., and Kim K., "Deep Abstraction and Weighted Feature Selection for Wi-Fi Impersonation Detection," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 3, pp. 621-636, 2018.

[3] Biesiada J. and Duch W., "Feature Selection for High Dimensional Data, A Pearson Redundancy Based Filter," *in Proceedings of Computer Recognition Systems*, Springer, pp. 242-249, 2007.

[4] Brankovic A., Hosseini M., and Piroddi L., "A Distributed Feature Selection Algorithm Based on Distance Correlation with an Application to Microarrays," *IEEE Transactions*, pp. 1-1, 2018.

[5] Breiman L. and Friedman J., "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, vol. 80, no. 391, pp. 580-598,1985.

[6] Chen R., Sun N., Chen X., Yang M., and Wu Q., "Supervised Feature Selection with a Stratified Feature Weighting Method," *IEEE Access*, vol. 6, pp. 15087-15098, 2018.

[7] Devarajan R. and Rao P., "An Efficient Intrusion Detection System by Using Behaviour Profiling and Statistical Approach Model," *The International Arab Journal of Information Technology*, vol. 18, no. 1, pp. 114-124, 2021.

[8] Haroun S., Seghir A., and Touati S., "Multiple Features Extraction and Selection for Detection and Classification of Stator Winding Faults," *IET Electric Power Applications*, vol. 12, no. 3, pp. 339-346, 2017.

[9] Heady., Luger G., Maccabe A., and Servilla M., "The Architecture of A Network Level Intrusion Detection System," Technical Report, University of New Mexico,1990.

[10] Hnaif A., Jaber K., Alia M., and Daghbosheh M., "Parallel Scalable Approximate Matching Algorithm for Network Intrusion Detection Systems," *The International Arab Journal of Information Technology*, vol. 18, no. 1, pp. 77-84, 2021.

[11] Lee J., Wangduk S., and Kim D., "Efficient Information-Theoretic Unsupervised Feature Selection," *Electronics Letters*, vol. 54, no. 2, pp.76-77, 2018.

[12] Liu Z., Huang J., Wang Y., and Cao D., "ECoFFeS: A Software Using Evolutionary Computation for Feature Selection in Drug Discovery," *IEEE Access*, vol. 6, pp. 20950-20963, 2018.

[13] Paxson V., "A System for Detecting Network Intruders in Real-Time," *in Proceedings of the 7th USENIX Security Symposium*, San Antonio, pp. 31-52, 1998.

[14] Roesch M., "Snort-lightweight Intrusion Detection for Networks," *in Proceedings of the 13th USENIX Conference on System Administration*, Seattle, pp. 229-238,1999.

[15] Toloueiashtian M., Golsorkhtabaramiri M., and Rad S., "Solving Point Coverage Problem in Wireless Sensor Networks Using Whale Optimization Algorithm," *The International Arab Journal of Information Technology*, vol. 18, no. 6, pp. 830-88, 2021.

**Suriya Prakash Jambunathan** is currently pursuing his Ph.D under Faculty of Information and Communication Engineering, Anna University Chennai, India. His Research center is Department of Electronics and Communication Engineering, M.Kumarasamy College of Engineering, Karur, Tamilnadu, India. He is guided by the Supervisor Dr. Suguna R and Joint Supervisor Dr. S. Palanivel Rajan .He completed his M.E in Computer Science and Engineering from St Peters University, Chennai. B.E in Computer Science and Engineering, from Raja College of Engineering and Technology, Madurai, Tamilnadu, India.

**Suguna Ramadass** is currently working as Professor in the Department of Computer Science and Engineering, at Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India. She received her Ph.D in Computer Science Engineering from Anna University Chennai, India, M.Tech in Computer Science and Engineering from IIT Madras ,Chennai, Tamilnadu, India, B.E in Computer Engineering, from Thiagarajar College of Engineering, Madurai, Tamilnadu, India. She has 28 years of academic experience.

**Palanivel Rajan Selva kumaran** is currently working as Associate Professor, Department of Electronics and Communication Engineering, M.Kumarasamy College of Engineering, Karur, Tamilnadu, India. He completed his Ph.D in Faculty of Information and Communication Engineering, Anna University Chennai, India.M.E in Communication Systems, Thiagarajar College of Engineering , Madurai, Tamilnadu, India. B.E in Electronics and Communication Engineering, from Raja College of Engineering and Technology, Madurai, Tamilnadu, India.