# Efficient Network Intrusion Detection Using PCA-Based Dimensionality Reduction of Features

## Abstract

The process of building a machine learning based Intrusion detection system for a dataset with high dimensions or features is complicated and takes a prolonged time to analyze. This research paper uses Principal Component Analysis to reduce the dimensions of the features and then uses various classifiers such as Random Forest (RF), Bayesian Network (BN), Linear and Quadratic Discriminant Analysis (LDA & QDA) for designing an IDS. The findings are classified into both binary class and multiple class. This research has used CICIDS2017 network intrusion dataset. With this dataset we have the feature dimensions reduced from 81 to 10. The results produced have an accuracy of 99.6%.

## Related works

In Sharafaldin et al., uses Random Forest Regressor for feature reduction and uses classifier such as KDD, Adaboost, RF, ID3 and Naïve Bayes. The highest precision value of 0.98 with RF and ID3. In Aksu et al. shows a Denial-of-Service IDS with Fischer Score algorithm. They achieved 99% accurate results. Marir et al., utilized a distributed Deep Belief Network as the dimensionality reduction approach. In Bansal et al. they have used data dimensionality reduction, where they achieved 98.93% accuracy with XGBoost as a classifier. Xia et al uses PCA for feature reduction and grey neural network as the classifier.

## Dataset Features

The CICIDS2017 intrusion detection and prevention dataset consists of five separate data files. Each file represents the network flow on each day totalling 5 days. The data on Monday represents benign and non-malicious data flow whereas the data from Tuesday to Friday showcases 14 different types of attacks in a non-uniform way. The dataset has total number of 2,830,108 records with 2,358,036 (83.3%) being benign records and remaining 471,454 (16.7%) being malicious attacks. Each record has 84 features or columns of data in which the last column represents the traffic status (benign/attack). The features are extracted with CICFlowMeter-V3 into a CSV file.

## Methodology and Experimental procedures

### Preprocessing

At first, a preprocessing function is applied to the dataset by mapping the IP address to an Integer presentation. The mapped IP includes the source and destination IP addresses converted to a number representation.

### Feature Reduction

PCA is a projection-based process that preserves the essential characteristics of the original dataset while projecting it into a subspace with less dimensions. The data is first preprocessed to normalise its variance and mean. The next step is to construct the covariance matrix, Eigen-vectors, and Eigen-values to create the reduction phase.

### Classification

An ML-based IDS framework is built and tested using a variety of classifiers, including RF, BN, LDA, and QDA. The framework has fewer features and evenly distributed class distributions. This study uses a 70:30 split of the data between training and testing sets.

### Multi-Class Combined Performance Metric

The overall accuracy of a classifier is used to assess its effectiveness. Unfortunately, in the presence of skewed data, this metric may fail to provide sufficient information. In this paper, we compare various binary and multi-class classification systems using the multi-class combined performance metric by combining four metrics (FAR, Accuracy, Detection Rate, and class distribution). The result will be a value between 1 to -1, where -1 is the worst and 1 being the best performing.

### Uniform Distribution based balancing (UDBB)

This section introduces the Uniform Distribution Based Balancing (UDBB) technique. UDBB is based on probability distributions learned and sampled [23]. In this technique, instances are sampled based on a learned distribution for each pair example of feature and class label. In order to re-sample new instances, a uniform distribution is considered.

## Result and Conclusion

All simulations are carried out in machine with Intel Core i7 with 3.3 GHz and 32GB Ram, running windows 10 OS. This research was successful in reducing the dimensions from 81 to 10 while maintaining a high accuracy. The highest accuracy achieved was 99.6% while using Random Forest as a classifier but Random Forest algorithm also took the highest time to build the model. The other algorithms used as a classifier also had similar results while the linear discriminant algorithm have varying accuracy with the difference in the number of features.