iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** V **Month of publication:** May 2019

DOI: https://doi.org/10.22214/ijraset.2019.5162

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

A Survey on Performance Analysis through Dimensional Reduction and Classification Algorithm using KDD Cup and UNSW-NB15 Dataset

Manu V[1], Suriya Prakash J[2], Charulatha M[3], Manjunatha M[4], Meghana K J[5]
[1, 3, 4, 5]*UG Student,* [2]*Assistant Professor, Dept. of CS Engg, Sapthagiri College of Engg, Visvesvaraya Technological Institute, Bengaluru, India*

Abstract: ***In this system the intrusion detection is one of the major research problems in network security. This is the process of monitoring and analyzing network traffic data to detect security violations. In this paper, we present the experimental results in our project to evaluate the different performance like (e.g., IDS, Malware, etc.). We analyze some different algorithms with dimensionality reduction and classification algorithm with the dataset that is constructed from the KDD CUP dataset. Data mining approach can also play a very important role in developing an intrusion and detection technique. The network traffic can be classified into normal and anomalous in order to detect intrusion detection. In our work, we use five (5) different algorithm's namely logistics regression, decision tree, random forest, KNN, KernelSVM are we used in the classification algorithm. The comparison of this classification algorithm is presented in this paper based upon their accuracy, timing, and performance to find out suitable algorithm's available and this method are performed in the spyder tool using UNSW-NB15 dataset.***
Keywords: Dimensional reduction, PCA, LDA, KernelPCA, KDD CUP Dataset, Classification algorithm's, logistics regression, Decision tree, Random forest, KNN, KernelSVM, UNSW-NB15 dataset.

I. INTRODUCTION

[1] Machine learning is a field of computer science that uses statistical techniques to give computer system the ability to "learn" with data, without being explicitly programmed. Machine Learning techniques are widely used in IDS due to its ability to classify normal/attack network packets by learning patterns based on the collected data. There are many results for classification of normal/attack; however, there is little work on classifying different attack types. In the modern world, the advanced internet technologies have made a huge collection of the data, which has become a major challenge for a human to analyze and processor to extract valuable information from the high dimensioned data. With the help of data mining techniques, this can be achieved easily. [2] Dimensionality reduction is a technique which uses feature selection and future extraction. In the feature, the selection is a technique which is used to find the good quality of relevant features from the original dataset using some objective measures. Nowadays, feature selection has become very big challenge issues in the field of [1] machine learning [2] Data mining. [3] Case-Based Reasoning. In feature extraction, the technique of extraction of features is used to get the most relevant information from the original data and to represent that information in a space of lower dimensionality. To select a new set of features, this technique is used. A linear or nonlinear combination of original features may be the transformation feature. The classification algorithm is the problem of identifying to which set of categories a new data belongs, on the basis of the training set and testing set. In our work this classification algorithm we use with principal component analysis (PCA) and without principal component analysis (PCA) of analysis. Using different algorithm's in our experimental work namely which as logistics regression, decision tree, random forest, KNN algorithm, kernel SVM.

II. PROBLEM STATEMENT

In this issue, different Dimensionality Reduction algorithms and Classification algorithms will be used to analyze the KDD CUP and USNW-NB15. Dataset to identifies better performance and accuracy. The data set will automatically be converted into a training set and test set based on the user input for performance measurement between different classifications algorithms.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

III. DIMENSIONALITY REDUCTION

Reduction of dimensionality is a series of machine learning techniques and statistics to reduce the number of random variables to be considered. It includes the selection of features and extraction of features. Reduction of dimensionality makes analyzing data much easier and faster without processing extraneous variables for machine learning algorithms, making machine learning algorithms, in turn, faster and simpler. Reduction of dimensions or reduction of dimensions is the process of reducing the number of random variables to be considered by obtaining a set of main variables. It can be divided into the extraction of selections and features. The project's scope is to conduct a comparative analysis of different algorithms to find the best accuracy.

IV. PROPOSED SYSTEM

A. Principal Component Analysis (PCA)

PCA is the most widely used linear reduction method. The PCA is a method of statistical data analysis that transforms the initial set of input variables into a different set of linear combinations, called the main components (PC). This PC contains specific variance properties. This reduces the system's dimensionality while retaining variable connection information.

B. Linear Discriminant Analysis (LDA)

LDA is a widely used reduction of dimensionality technique. In some dataset experiments, quantity growth is greater in existing cases where dimensions are greater or fewer characteristics and the occurrence of characteristics is significantly greater than the sample size. LDA creates a linear combination that yields the greater mean differences between the classes described. LDA's main goal is to maximize measurements between classes while minimizing measurements within the class.

C. Kernel Principal Component Analysis (KernelPCA)

Various different approaches along with kernel functions were also studied as extensions to the PCA to solve the non-linearity problem. Before performing PCA, the kernel PCA maps the samples into high-dimensional kernel space to convert the nonlinear distribution of input data to linear distribution. SPCA's basic principle is to transform original input vectors into a high-dimensional F-space feature with a nonlinear function and then calculate the linear PCA in feature space.

V. SEQUENCE ARCHITECTURE

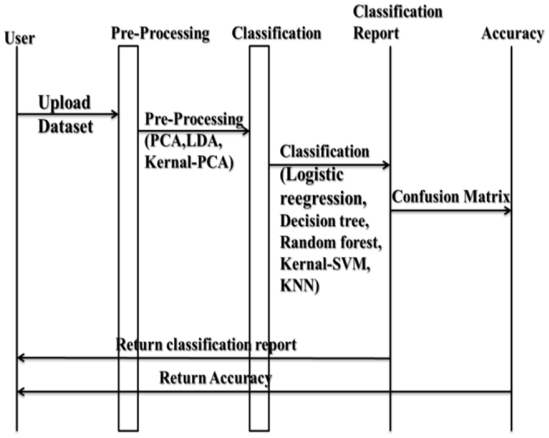


Figure 1: Sequence diagram

A. Upload Dataset

We will create a new dataset and upload the dataset at a particular file path to fetch and perform the given task when it is needed using [.CSV] file.

B. Labeled Training Dataset

High-quality labeled training datasets are usually difficult and expensive to produce for supervised and semi-supervised machine learning algorithms due to a large amount of time required to label the data. Although they do not need to be labeled, it can also be difficult and costly to produce high-quality datasets for unsupervised learning.

C. *Unlabelled Testing Dataset*

Unlabelled data typically consists of samples of natural or man-made artifacts that you can get from the world relatively easily. Some unlabelled data examples could include photos, audio recordings, videos, news articles, tweets, x-rays.

D. *Pre-processing*

The transformations that are applied to our data before the algorithm is fed. Data Preprocessing is a method of converting raw data into a clean set of data. In other words, it is collected in raw format whenever the data is collected from different sources, which is not feasible for analysis.

E. *Dimensionality Reduction [PCA]*

Principal Component Analysis (PCA) is the classical statistical technique which is widely used to reduce the dimensionality of the given dataset consisting of an enormous amount of interrelated variables. PCA is mainly used to reduce the dimensionality by transforming the original dataset into a new set of variables called principal components, in which largest variance present in the original dataset is captured by the highest component in order to extract the most important data or information.

F. *Classification Algorithm*

An algorithm implementing classification is known as a classifier, especially in a concrete implementation. Sometimes the term "classifier" also refers to the mathematical function, implemented by a classification algorithm, which maps data input into a category. There is quite a variety of terminology across fields.

G. *Accuracy*

In machine learning, a number of metrics are used to measure a model's predictive accuracy. The choice of precision metrics depends on the task of learning the machine. These metrics should be reviewed to determine if your model performs well.

H. *Confusion Matrix*

A confusion table (sometimes also called a confusion matrix) in predictive analytics is a table with two rows and two columns reporting the number of false positives, false negatives, true positives, and true negatives. This allows for a more detailed analysis than just the proportion of correct classifications (precision).

VI. DATASET

KDD CUP Dataset was built on a network intrusion detector, a predictive model should be capable of distinguishing between “bad’” connections, called intrusions or attacks, and “good” normal connections.

UNSW-NB15 — The UNSW-NB 15 dataset raw network packets were created by the IXIA Perfect Storm tool at the Australian Cyber Security Centre (ACCS) Cyber Range Lab to generate a hybrid of real modern normal activity and contemporary synthetic attack behaviors.

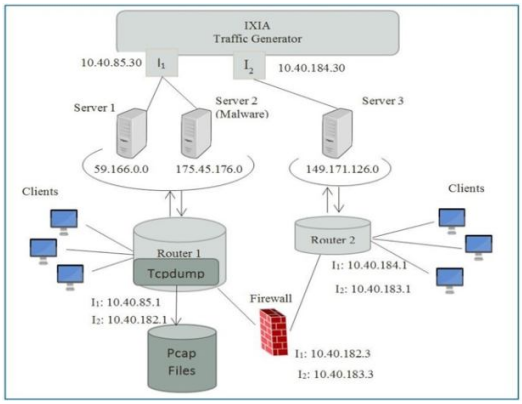


Figure2: Dataset diagram

TCP dump tool is utilized to capture 100 GB of the raw traffic. The Argus, Bro-IDS tools are used and twelve algorithms are developed to generate totally 49 features with the class label.

VII. ALGORITHM'S USED

A. PCA algorithm

Principal Component Analysis is one of the most widely used techniques for data analysis and compression dimensionality reduction. It is based on converting a relatively large number of variables into a smaller number of uncorrelated variables by finding a few linear orthogonal combinations of the original variables with the greatest variance.PCA reduces the number of dimensions needed to classify new data and produces a set of main components that are pairs of orthonormal self-value / eigenvector. The main component analysis steps are outlined below.

1) Algorithm

```
from sklearn.decomposition import PCA
pca = PCA(n_components = 100)
X_train = pca.fit_transform(X_train)
X_test = pca.transform(X_test)
explained_variance= pca.explained_variance_ratio_
```

B. Logistics Regression

Logistic Regression is a classification algorithm for machine learning that is used to predict a categorical dependent variable's probability. The dependent variable in logistic regression is a binary variable containing data coded as either 1 (yes, success, etc.) or 0 (no, failure, etc.).

1) Algorithm

```
fromsklearn.linear_model  import
LogisticRegression
classifier =
LogisticRegression(random_state = 0)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
```

C. Decision Tree

Decision Trees can be used as models for classification or regression. A tree structure is built that breaks down the dataset into smaller subsets leading to a prediction eventually.

1) Algorithm

```
from sklearn. tree import Decision Tree Classifier
classifier=DecisionTreeClassifier(criterion=
'entropy',random_state=0)
classifier.fit(X_train,y_train)
y_pred = classifier.predict(X_test)
```

D. Random Forest

Random forest algorithm is a monitored algorithm for classification. As the name suggests, with a number of trees, this algorithm creates the forest.

The missing values will be handled by the random forest classifier. If we have more trees in the forest, the model will not be overfitted by random forest classifier.

Random forests create decision trees on randomly selected data samples, get a prediction from each tree, and by voting select the best solution. It also provides a good indicator of the significance of the feature. Random forests have a variety of applications, including recommendation engines, classification of images, and selection of features.

1) Algorithm

```
from sklearn.ensemble import Random Forest Classifier
classifier=RandomForestClassifier(n_estimators=10,criterion='entropy',random_state=0)
classifier.fit(X_train,y_train)
y_pred = classifier.predict(X_test)
```

E. KNN Algorithm

the k-nearest neighbor algorithm is a nonparametric method used for classification and regression. In both cases, the input consists of k closest training example in the feature space. The output depends on whether KNN is used for classification or regression. KNN is typical instance-based learning whether the function is only approximated locally and all computation is deferred until classification. The KNN algorithm is among the simplest of all machine learning algorithm. Both the classification and regression useful technique can be used to assign a weight to the contribution of the neighbors so that the nearer neighbors contribute more to the average than the more distant ones. The neighbors are taken from a set of objects from the class (for KNN classification). A peculiarity of the KNN algorithm is that it is sensitive to the local structure of the data.

1) Algorithm

```
from sklearn.neighbors import KNeighbors Classifier
classifier=KNeighborsClassifier(n_neighbors=5,metric='minkowski',p=2)
classifier.fit(X_train,y_train)
y_pred = classifier.predict(X_test)
```

F. KernelSVM

Kernel methods are a class of algorithms for pattern analysis in machine learning, with the support vector machine (SVM) being the best-known member. The general pattern analysis task is to find and study general types of relationships in datasets (e.g. clusters, rankings, main components, correlations, classifications). Functions for the SVM kernel. SVM algorithms use a set of kernel-defined mathematical functions. The kernel's function is to take input data and transform it into the form required. Various SVM algorithms use various kernel function types.

1) Algorithm

```
from sklearn.svm import SVC
classifier=SVC(kernel='rbf',random_state=0)
classifier.fit(X_train,y_train)
y_pred = classifier.predict(X_test)
```

VIII. LITERATURE SURVEY

In [1] this study, we used the Random Forest algorithm, an efficient supervised machine learning algorithm for IDS, to analyze class-specific detection of Kyoto 2006 + datasets. The original 3 classes (i.e., normal, known attack unknown attack) were first refined into 6 classes (i.e., normal, unknown, shellcode, IDS shellcode, malware, IDS). We then built a test dataset and the two training data. Next, we built a test dataset and two training datasets that vary in size between classes to assess the performance of detecting different types of attack. Although we obtained a high overall detection performance when trained with the first training set (0.99 of accuracy, recall, F1-score, and F2-score), we found that performance differs greatly for each class (as low as 0.16 of F1-score for shellcode attack). That's why we built the second training set using random under-sampling to set the size of the whole class equal to the number of instances of the smallest class (i.e., shellcode). The evaluation resulted in much lower performance, which was disappointing for all classes. We believe that data size was not enough, and training with the same size class may not be ideal for the approaches to machine learning. We also note that the unknown attack class still shows a good performance, 0.90 F1-score, suggesting that the unknown attack has a distinct pattern.

The [2] purpose of this experimental work was to find out which NIDS (network intrusion detection system) was the best available classification technique. This study is carried out by analyzing the NSL-KDD dataset and the performance of classification algorithms is observed. The study shows that in classifying the intrusions, decision trees classifiers are best. With respect to the accuracy, specificity, and sensitivity, Random Forest has outperformed, while IBK consumes less time compared to others. The main goal was to achieve a better rate of detection of intrusion, to lower the rate of false negatives. This work can be extended by combining various data mining algorithms with performance-enhancing data reduction techniques. In identifying new and unusual attacks, an intrusion detection system based on hybrid classification techniques would be quick and robust.

In [3] a survey is made on major challenges and issues in dimensionality reduction. If the dimensionality of dataset increased, then the volume of the space increases so fast that the available data becomes sparse. Usually, a larger percentage of the training data resides in the corners of the feature which is more difficult to classify. Hence high dimensionality leads to a problem known as "Curse of Dimensionality" that specifically makes it difficult to perform classification on a dataset having a large

number of dimensions. Dimensionality reduction can be used for downsizing the input data i.e., more relevant for further analysis. The reduced dataset contains variance from a large dataset and without any loss of important features. It has also made easy to detect and us from real word data. PCA is the most popularly used linear dimensionality reduction technique. The linear method can work only with linear data and not work with real data efficiently because of complexity and high-dimensionality. PCA can work with structured and steady dataset. PCA is a statistical data analysis method the=at transforms the initial set of input variables into a various set of linear combinations, called as the Principal Component (PC). This PC contains specific properties with respect to variances, which helps to reduce the dimensionality of the system while retaining information on the variable connections. LDA has an issue with lack of sample data per class does degrade the classification performance as significantly due to the generalization of decision for arbitrary data with noise regulation. Robustness improvement is pursued as the other critical issue in LDA for better classification performance in a noisy environment. The main aim of LDA is to maximize the between-class measure while minimizing within-class measure. Kernel-PCA is used to solve the problem of non-linearity, various different approaches are used along with kernel functions which also been studied as an extension to the PCA. Kernel PCA is used to transform original input vectors to a high dimensional feature space with nonlinear function and to calculate the linear PCA in feature space. Kernel PCA computes principal Eigenvector for kernel matrix rather than covariance matrix. Kernel PCA has been applied to successfully to different domains face recognition, speech recognition, novelty detection, etc. comparing all the three techniques from this paper, we can conclude that combination of the method may also be used to overcome the disadvantages of one method over another.

In [4] we learned about PCA and LDA of dimensionality reduction techniques. In this paper dimensionality reduction is defined as the processes of projecting high-dimensional data to much lower-dimensional space. Dimensional reduction methods variously applied in the regression, classification, feature analysis and visualization. PCA is the linear method which is used to perform a dimensionality reduction by embedding the data into a linear dimensional. PCA is the widest unsupervised linear method. The result of PCA is the lower dimension representation from original data that describe as much of the variance in the data. This can be reached by finding the linear basis of reduced dimensionality for data, in which the amount of variance in the data is maximal. PCA and classical scaling suffer from two main drawbacks. First, in PCA, the size of the covariance matrix is proportional to the dimensionality of the data-points. Second, the cost function reveals that PCA and Classical scaling focus mainly on retaining large pairwise distances, instead of focusing on retaining the small pairwise distances, which is actually more important. LDA is a method to find a linear transformation that maximizes class separability in the reduced dimensional space. The criteria in LDA is to maximize between-class scatter and minimize within-class scatter.

The [5] approach proposed improved the speed of detection. Selection of features reduced the total number of data set features (32 basic features and 116 derived features). This reduction means that less data is needed to train the classifier due to the smaller search space. Paper reports a new approach to the CBID that can produce better and more accurate results by identifying the attack category rather than the exact type of attack. This result also indicates that feature selection analytical solutions are not based on the trial and error. An important goal in the reported work is the possibility and feasibility of detecting intrusions based on characterizing various types of attacks such as DoS, probes, U2R and R2L attacks. It seems that the results of this investigation are promising. Results indicate that a small number of carefully selected network features can be used to identify the normal state of the network and attack category. On the other hand, it is proved that the detection of intrusion is not connected with certain features. Experimental results show that dimensional reduction and identification of effective network features for category-based selection can reduce process time in an intrusion detection system while maintaining accuracy within an acceptable range of detection. The PCA method is used to determine an optimal set of features to speed up the detection process. Experimental results show that feature reduction can improve detection rates for the category-based detection approach while maintaining detection accuracy within an acceptable range. KNN classification method is used in this paper to classify the attacks. Experimental results show that feature reduction will significantly speed up the intrusion attempts train and testing periods.

This [6] paper uses classification algorithms J48, Naïve Bayes, LTM, REP, Decision table, K-Star, Simple Logistics, Iterative Classifier, IBK, and Filtered Classifier to carry out a comparative study to predict breast cancer. The datasets are taken from the Wisconsin breast cancer datasets of 10 attributes with 286 instances. From the results, it was observed that Naïve Bayes, K-Star, IBK, and Filtered Classifier performs well with regard to accuracy J48 and Filtered classifier and execution time is 0 sec. So we can conclude that with 76 percent accuracy and 0 sec execution time, Filtered Classifier is the best. By considering more attributes, applying some dimensionality reduction algorithms and other supervised as well as unsupervised methods, we will compare results in the future and compare their performance.

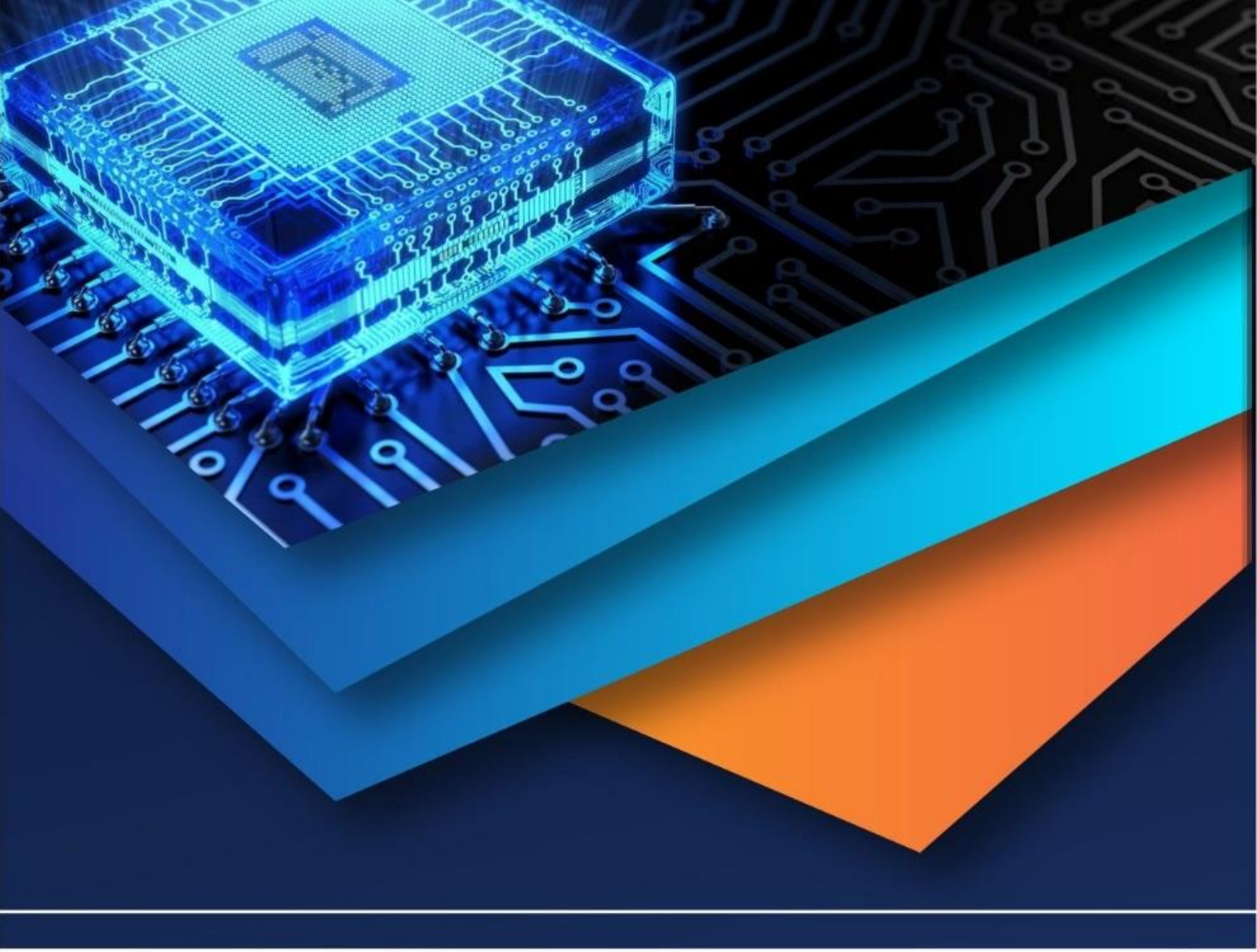
In [7] this paper, we propose a novel method of classifying network intrusion detection from the most renowned KDD cup dataset using ensemble learning scheme. We have shown that the most accurate detection is provided by reducing the dimensionality of the large dataset. In addition, for a proper comparison, several machine learning algorithms are used to generate accuracy metrics and further analyzed. Our approach found that all other learning techniques were outperformed by this algorithm. Our goal is to analyze the intrusion data of the network and find the best components and use them for the analysis of the attack. This scheme can be used to increase its prediction performance for future data packets in parallel with the intrusion detection system. Empirical results show that the reduction in input dimensionality can provide a lightweight intrusion detection system that can be embedded with the vulnerable system to generate correct classification with an improvement in execution time of significance. In the previous sections, while classifying the dataset using several well-known machine learning algorithms, we tried to present different scenarios. If we can adjust some key parameters, a single learning algorithm can produce significant improvements in classification. We analyzed those details and suggested the best configuration to use when solving this particular problem type. We will use evolutionary algorithms in our future work to further accelerate the speed and accuracy of classification. In addition, we have the plan to implement an online NIDS that can provide real-time feedback to the system so that the offline detection method can eradicate the unintentional delay.

IX. CONCLUSION

In our project, we conclude by identifying the best-fit algorithm for KDD CUP dataset as well as 10 percent KDD CUP and UNSW-NB15dataset by means of the reduction algorithm for dimensionality. Using the dimensionality reduction algorithm, classification algorithm and performance measurement, the KDD cup can be observed to use test data and training data. For each paper being surveyed, the pros and cons of the existing system are identified. It recognizes the need for a more accurate working system. How to look at the current system's lack of semantic analysis. This overview can also help researchers and analysts build a more sophisticated system. To find good accuracy with which algorithms are the best fit.

REFERENCES

[1] Kinam Park, Youngrok Song, Yun-Gyung Cheong, Classification of Attack Types for Intrusion Detection Systems using a Machine Learning Algorithm, DOI 10.1109/ Big Data Service.2018.00050, IEEE paper 2018.
[2] Himadri Chauhan, Vipin Kumar, Sumit Pundir and Emmanuel S. Pilli, A Comparative Study of Classification Techniques for Intrusion Detection, DOI 10.1109/ ISCBI 2013.16, IEEE paper 2018
[3] Arun Kumar Venugopal, "A Survey on Dimensionality Reduction Technique" International journal of emerging of trends and & technology in computer science (IJETTCS),site:www.ijettcs.org.Email:editor@ijettcs.org,volume 3, Issue,November-December 2014,PP :ISSN 2278-6856
[4] Laurens van der maaten Tilburg centre for Creative Computing", Dimensionality Reduction:" A Compartive ReviewTilburg,University,http://www.uvt.nl/ticc, FO.Box90153,Email:ticc@uvt.nl October 26, 2009, TiCC TR 2009-005.
[5] Gholam Reza Zargar, Tania Baghaie, "Category-Based Intrusion Detection UsingPCA", Email:zargar@vu.iust.ac.ir,baghaie@vu.iust.ac.ir,2012.
[6] Dr. S. Senthil1 Deepa B.G2 Aishwarya B3, "Comparison of Classification Algorithms for Predicting Breast Cancer", International Journal for Scientific Research & Development| Vol. 4, Issue 12, 2017 | ISSN (online): 2321-0613.
[7] Deeman Yousif Mahmood, Classification Trees with Logistic Regression Functions for Network-Based Intrusion Detection System, Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 3, Ver. IV (May - June 2017)
[8] A Review on Dimensionality Reduction Techniques in Data Mining, Author: Prof. Sumit Sharma, Computer Engineering and Intelligent Systems, PP: ISSN 2222-1719(paper)ISSN 2222-2863(online), volume9, No.1, 2018
[9] An Actual Survey of Dimensionality Reduction, Author: Alireza Sarveniazi, American Journal of Computational Mathematics, 2014, 4, 55-72, Published Online March 2014 in SciRes.http://www.scirp.org/journal/ajcm.
[10] SSENet-2011: A Network Intrusion Detection System Dataset and its Comparison with KDD CUP Dataset, Author: Vasudevan A.R, Dept of CSE, National Institute of Technology Tiruchirappalli, Tamil Nadu State, Email: vasudevan@nitt.edu.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY