International Journal of Electrical, Electronics and Computer Engineering **7**(2): 69-76(2018)

ISSN No. (Online): 2277-2626

Intrusion Detection Framework using Correlation based Feature Selection over Cloud

Sanjeet Choudhary [1], Varsha Namdeo [2] and Abhijit Dwivedi [3]

[1]M. Tech. Scholar, Department of Computer Science & Engineering, RKDF Institute of Science & Technology, Bhopal (Madhya Pradesh), India

[2]Associate Professor & Head, Department of Computer Science & Engineering, RKDF Institute of Science & Technology, Bhopal (Madhya Pradesh), India

[3]Assistant Professor, Department of Computer Science & Engineering, RKDF Institute of Science & Technology, Bhopal (Madhya Pradesh), India

(Corresponding author: Sanjeet Choudhary)

(Received 07 September, 2018 Accepted 06 November, 2018)

(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Network Intrusion Detection (NID) has been considered to be one of the most promising methods for defending complex and dynamic intrusion behaviours. Addressing the need of computational intensive tasks of building intrusion detection system, cloud computing (CC) paradigm is evolved. Such CC platforms ease the complex tasks of Intrusion Detection using ML by offering scalable infrastructure at low cost. It is often observed that selecting a small number of predictive attributes over a large number of attributes, if properly used in the combination, are fully predictive of the class label. Apart from presenting the brief survey in the field of IDS, the major contribution of this paper is to present the most appropriate attribute selection algorithm. The correlation based feature selection method filters the features that are most correlated with target vector. The paper also proposes the framework for NID based on proposed correlation based feature selection technique. The framework is tested to predict the instances of famous KDD CUP 99 dataset. In this paper, an attempt is being made to show that how feature selection techniques affect the classification tasks. The proposed work is implemented and evaluated over Microsoft's Azure Machine Learning platform, which is the most prominent part of paper. The efficacy and superiority of classification framework is tested and validated against benchmark classifiers on basis of detection rate, accuracy, false positives and precision.

Keywords: Computer Vision, Computer Networks, Cloud Computing, Data Science, Intrusion Detection System, Feature Selection, Data Classification, Data Mining, Machine Learning, Microsoft Azure.

I. INTRODUCTION

The innovations in technology have led to closer access of information over the network. Technology has effortlessly improved the access of data over the network for the organizations and users. However, on the other hand, it has also exposed the network with the various kinds of threats and intrusions.

Intrusion Detection Systems (IDSs) intend at identifying attacks against computer systems and networks or, in general, against data. Training, testing and evaluation of IDS with real network traffic is significant challenge [1]. In fact, it is difficult to provide efficient IDS and to maintain them in such a secure state during their lifetime and utilization. Figure 1 shows the IDS in a typical Network. In the recent years NID techniques and Machine Learning (ML) have attracted the researchers much. ML algorithms can learn from data, to detect irregularities and special patterns automatically, to make predictions and future decisions. Due to the incessant growth in volume and variety of data, ML becomes cumbersome. This means applying statistical and analytical models to data will take more effort in terms of resources such as CPU and Memory. In that case, feature selection which is considered as an important pre-processing step aids in building IDS with less effort.

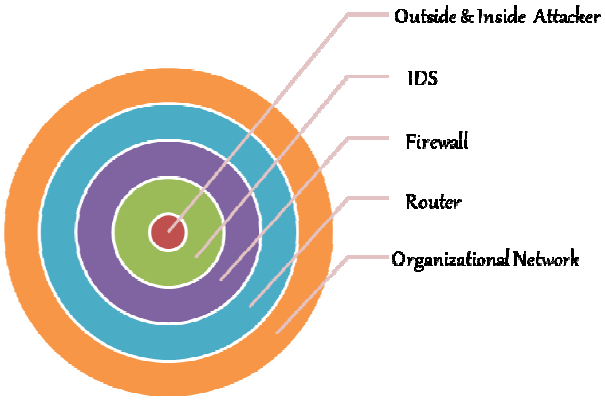

Fig. 1. NIDS in Network.

Feature selection is a method of choosing the optimal (relevant) feature subset from the original set of features. It also reduces the dimensionality of the data and allows ML algorithms to run faster and more effectively. Accuracy on future classification can also be improved in some cases. Also the memory requirements of algorithms are relaxed and results are more compact and interpreted as representation of the target concept.

It is divided into three categories: the filter method [2, 3], the wrapper method [4, 5] and the embedded method. The filter method selects the most use optimal features and is independent of model types.

The well known KDD Cup 99 dataset [6] is a typical example of large-scale intrusion evaluation dataset. Addressing the storage and computational requirements for intrusion detection tasks, cloud computing (CC) paradigm is used. One of the major benefits of CC is scalability that turns out to be a valuable option for expedite the DM and Intelligence task [7]. The popular statistical tools and environments like Octave, R and Python are now part of cloud platforms.

In this work, architectural framework for IDS is proposed which is based on correlation based feature selection. The architecture is implemented over one of the famous cloud platform. An attempt is being made to classify the attacks and normal transactions. It is studied that how feature selection techniques affect the classification tasks. Also, the classification is carried out with 2 benchmark ML algorithms in literature. To evaluate the efficiency of IDS, the major parameters are: Detection Rate, Accuracy, and False Positives. The proposed work is evaluated on these popular and effective metrics.

In Section 2 a brief review, on attack types and most relevant papers in the domain, is presented. Section 3 presents the Correlation Based Feature Selection Algorithm and Proposed Intrusion Detection Framework with correlation based feature selection algorithm. Experiment Setup for proposed Work, Dataset description, Results and Analysis is presented in Section 4. The proposed work is evaluated on the basis of various parameters. The comparison against benchmark models is also presented in this section. Finally the paper is concluded in section 5 with remarks for future work.

II. LITERATURE REVIEW

In general, intrusion can be considered deviation from normal expected use of the system. Intrusion detection has many common challenges as of fraud detection and fault management/localization. One of the widely used taxonomy divides attacks into four classes: Denial of Service (**DoS**), **Probe**, User to Root (**U2R**) and Remote to Local (**R2L**) [8, 9, 10]. The class **Normal** represents network traffic which is considered attack-free, is used later in this work.

A framework for IDS is proposed in [8]. Synthetic Minority Oversampling Technique (SMOTE) is applied to deal with class imbalance in the training dataset. To reduce the feature set a method based on Information Gain is applied before applying Random Forest classifier. Empirical results show that framework [8] gives better performance in designing IDS that is efficient and effective for NIDS.

Author [9] contributed by determining the appropriate feature selection algorithm for selecting the relevant features from 41 features. They used statistical techniques to classify the instances in data. To validate the efficacy of proposed algorithm against other state-of-the-art methods, the analysis of results for various models is done by comparing their accuracy, detection rate, FAR etc. Taking inspiration from IDS that make use of ML potential to improve accuracy in detecting anomalies, the paper [11] proposes that cloud based ML can be used in order to detect and classify the packet data by feeding it into a cloud based ML web service. Azure Machine Learning Studio [16, 17] used for deploying the proposed model.

A new RPFMI (redundant penalty between features based on Mutual Information) is proposed in [12] with the ability to select optimal features. Three factors are considered in this new algorithm: the redundancy between features, the impact between selected features and classes and the relationship between candidate features and classes.

Subba *et al.* [13] proposed an algorithm using Logistic Regression (LR) and Linear Discriminant Analysis (LDA). The authors have reduced the features to 23 using suitable feature selection.

Due to their higher computational efficiency, the IDSs based on LR were found fit for operation in real-time networks. The work [13] successfully achieved the accuracy up to 95.44%.

Davaraju *et al.* [14] made an effort to improve the performance of model for IDS using Neural Network (NN) IDS dataset and achieved the accuracy up to 96.33%. The NN models used in this work are Probabilistic, Feed Forward and Radial Basis etc. The minimum number of optimal features was reduced to 9.

Prwez and Chatterjee [15] proposed a model for IDS based on AdaBoost (Boosting) method which is an ensemble method. The ensemble methods uses one of the classifiers as base classifier. Here for IDS, they used Decision Tree of height 1. The purpose of using decision stumps is to build a strong classifier by merging the weak classifiers.

A Graphical ML Workspace is an extraordinary feature of Microsoft Azure ML Studio (MAMLS) [16]. The major parts of it are (a) ML studio, (b) ML Gallery and (c) ML Web Service Management. Azure ML studio comprise of ML processes from start to end.

III. PROPOSED WORK

A. Proposed Classification and Prediction Framework

The Framework proposed is presented in Figure 2, which is divided in five major modules. The task of five modules is broken to fulfill the objective of intrusion detection. The five modules are mentioned below:

1. Data Collection Module: The module collects and supplies the traffic data which is considered as training set. For evaluating the Framework, Online Traffic data collection is also done through this sub module. The data is supplied to Data Processing module for pre-processing.

2. Data Processing (DP) Module: This module comprise of: Data Normalization, Transformation and Proposed Feature Selection which is treated as part of DP Module is applied to determine attributes or features that are more predictive for the output.

3. Model Building Module: This module Train, Build and Score the model built for IDS.

4. Test Module: In test module, the inferences from DP module is called and applied to online traffic data and resultant data is sent to attack recognition phase which works on the model built in previous phase (i.e. phase 3). The 'Attack Recognition Phase' predicts the intrusion and generates an alarm in case of any attack.

5. Evaluation Module: In this module, evaluation of IDS will be done on various parameters: Accuracy, False Positives and Detection Rate etc.

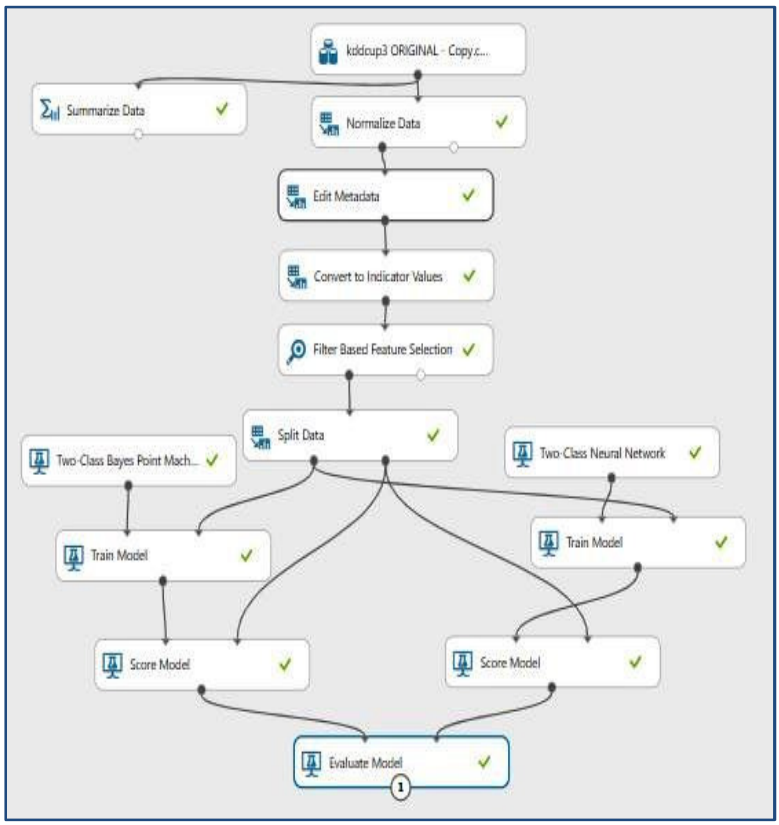

Fig. 2. Proposed Classification and Predictive Framework.

B. Correlation Based Feature Selection Algorithm

Correlation coefficient is value say **r** value. For any two columns or labels, it returns **r** that indicates the strength of the correlation between two labels (here in output class and any column). The value of **r** is not affected by changes of scale of values of two labels. Here, this work used Pearson Correlation coefficient

Pearson's correlation coefficient is computed by taking the covariance of two variables and dividing by the product of their standard deviations.

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Here Y is class label and X is attribute ranging from 1 to 41.

Algorithm 3. Proposed algorithm for Feature Selection (Filter)

```
INPUT: Attribute vector (m), Class vector, number of attributes to select (n)
OUTPUT: A – Attributes A[i...n]
begin
Declare Arrays F[m], A[n] of type float
for i = 1 to m
F[i] = compute_R(A[i] , Class)
end for
sort(F)
for i = 1 to n
A[i] = F[i]
end for
print(A)
end

compute_R(A, Class)                    //Method to Compute R
{
float R
Calculate covariance(v) and
Calculate standard_deviation(d)
R = v/d                                // Calculate Correlation (R)
return R
}
```

The return value of R is lies between -1 and 1, where: 1, -1, 0 indicates a strong positive, strong negative and zero relationship respectively. The algorithm determined 7 features that are strongly correlated with class label are set for algorithm.

IV. EXPERIMENT, RESULT ANALYSIS AND DISCUSSION

A. Experiment

The actual experimental model is shown in Figure 2. The implementation of proposed architecture comprising correlation based Feature selection algorithm is done at MAMLS [16-17].

KDD CUP 1999 Data (KDD99) [6] is the dataset used in the evaluate ML technique. The full KDD99 dataset Contain 4,898,431 records and each record contain 41 features [8-10]. Although there is no bound on computing power over CC environment, to reduce computing time we use the 10% portion use of dataset. The details of attack categories and number of instances of specific types are shown in Table 1. The four categories are mentioned in [8-10]:

Table 1: KDD Cup 1999 Dataset Description.

Dataset: 10-percent KDD		Total Attacks	Total Instances
DoS	391458	**396743**	**494021**
U2R	52		
R2L	1126		
Probe	4107		
Normal	**97278**		

B. Results

The number of features selected using correlation based feature selection is actually 7 which are much lesser than the mutual information method mentioned in [9]. The performance results obtained with proposed and benchmark models are shown below in Table 2.

Table 2: Performance Comparison.

Models	TPR	FP	Accuracy	Precision
Naive Bayes [9]	0.92		92.73	0.98
Neural Network	0.99	1542	99.3	0.97
Naive Bayes (over Cloud)	0.98	2790	98.5	0.94
Proposed Model	0.99	346	99.8	0.99

The results, on the basis of important performance metrics, are graphically represented in Figure 3 to 6.

Accuracy. Accuracy is defined as the number of correctly classified instances divided by the total number of instances:

$$Accuracy = \frac{Number\ of\ correct\ Predictions}{Number\ of\ Instances}$$

The results obtained are shown in Graph figure 3.

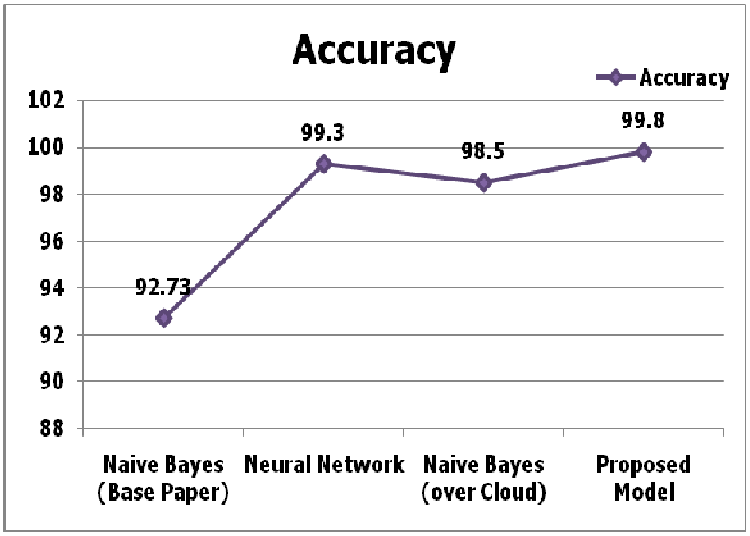

Fig. 3. Comparison of Accuracy.

False Positive. The number of positives that are wrongly classified as negatives i.e. Numbers of normal instances that are wrongly classified as attack are known as false positives. The results obtained are shown in Graph figure 4.

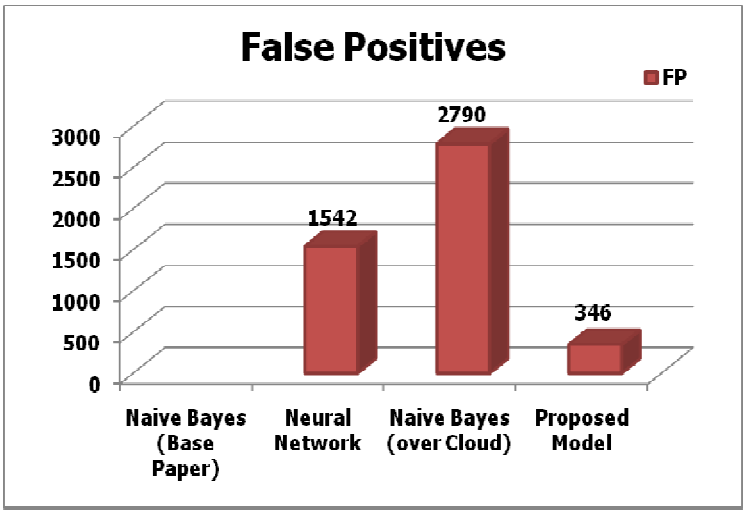

Fig. 4. Comparison of False Positive.

True Positive Rate. This is also called Detection Rate, and represented using formula

$$True\ Positive\ Rate = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

The results obtained are shown in Graph figure 5

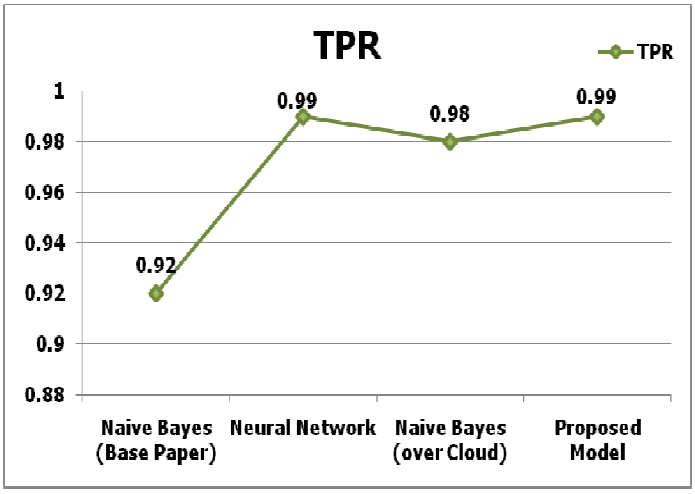

Fig. 5. Comparison of True Positive Rate.

Precision. Precision (PR) is the portion of predicted positive values which actually turn out to be positive. When FP = 0 the precision reach a value of 1 i.e. FP Rate is least. Precision is represented using formula

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

The results obtained are shown in Graph figure 6.

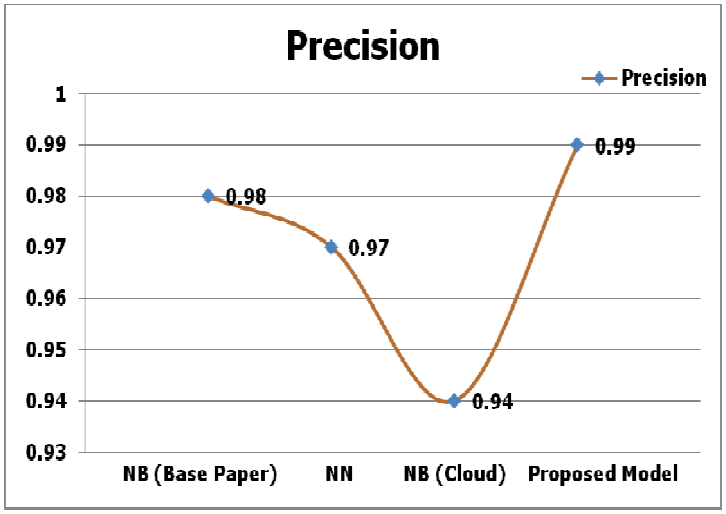

Fig. 6. Comparison of Precision.

C. Discussion

The results attained by proposed framework are much better in comparison to benchmark models and as attained by model in base paper [9]. Also, the False Positives are only few.

The overall accuracy attained using proposed framework with correlation based feature selection is above 99% as shown while it is 92.73% in base paper. Also the results are evaluated against benchmark models which are also higher due to better and faster algorithm conversion in cloud environment.

V. CONCLUSION

The power of standalone user system has lot of limitations in terms of processing and memory. Also, building the predictive models based on ML techniques is not simple. Extensive ML platforms with several components are now provided by several companies for computational intensive tasks.

The key contributions of this paper include survey of relevant work in the field, a proposal of efficient correlation based feature selection technique to improve the performance of classifier for IDS and a architectural framework based on proposed algorithm. Apart from this, the proposed method selects minimum number of features. The proposed work is **implemented** and **evaluate**d over MAMLS which is computationally efficient with no resource limitations. To show the efficacy, the proposed framework is compared with benchmark classifiers on basis of suitable parameters.

The proposed framework can prove potential in addressing multi-class classification and prediction problems. So, further research in the proposed field will motivate to test and evaluate the model with imbalanced datasets, different optimization techniques and parameters.

REFERENCES

[1]. U. S. K. P. M. Thanthrige, J. Samarabandu and X. Wang, (2016). "Machine learning techniques for intrusion detection on public dataset," 2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Vancouver, BC, 2016, pp. 1-4.

[2]. Peng, H.C.; Long, F.H.; Ding, C. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th *International Conference on Machine Learning, Washington,* DC, USA, 21–24 August 2003.

[3]. Mohamed, N.S.; Zainudin, S.; Othman, Z.A. (2017). Metaheuristic approach for an enhanced mRMR filter method for classification using drug response microarray data. *Expert Syst. Appl.,* **90**, 224–231.

[4]. Kohavi, R.; John, G. (1997). Wrappers for feature subset selection. *Artif. Intell.*, **97**, 273–324.

[5]. Hui, K.H.; Ooi, C.S.; Lim, M.H.; Leong, M.S.; Al-Obaidi, S.M. (2017). An improved wrapper-based feature selection method for machinery fault diagnosis. *PLoS ONE* 2017, **12**, e0189143.

[6]. KDD CUP (1999). Availabe on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html October 2007

[7]. L. Chih-Wei, H.Chih-Ming, C.Chih-Hung, Y.Chao-Tung, (2013). "An Improvement to Data Service in Cloud Computing with Content Sensitive Transaction Analysis and Adaptation, Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual, 2013, pp.463–468.

[8]. A. Tesfahun and D. L. Bhaskari, (2013). "Intrusion Detection Using Random Forests Classifier with SMOTE and Feature Reduction," 2013 *International Conference on Cloud & Ubiquitous Computing & Emerging Technologies*, Pune, pp. 127-132.

[9]. P. Kushwaha, H. Buckchash and B. Raman, (2017). "Anomaly based intrusion detection using filter based feature selection on KDD-CUP 99," TENCON 2017 - 2017 *IEEE Region 10 Conference, Penang*, pp. 839-844.

[10]. A. D. Landress, (2016). "A hybrid approach to reducing the false positive rate in unsupervised machine learning intrusion detection," SoutheastCon 2016, Norfolk, VA, 2016, pp. 1-6.

[11]. S. Miller, K. Curran and T. Lunney, (2016). "Cloud-based machine learning for the detection of anonymous web proxies," 2016 27th Irish Signals and Systems Conference (ISSC), Londonderry, pp. 1-6.

[12]. Zhao, F.; Zhao, J.; Niu, X.; Luo, S.; Xin, Y. A Filter (2018). Feature Selection Algorithm Based on Mutual Information for Intrusion Detection. *Appl. Sci.*, **8**, 1535.

[13]. B. Subba, S. Biswas and S. Karmakar, (2015). "Intrusion detection systems using linear discriminant analysis and logistic regression," 2015 Annual IEEE India Conference (INDICON), New Delhi, pp. 1-6.

[14]. S. Devaraju and S. Ramakrishnan, (2014). "Performance comparison for intrusion detection system using neural network with KDD dataset," *ICTACT Journal on Soft Computing,* vol. **4**, no. 3.

[15]. M. T. Prwez and K. Chatterjee, (2016). "A Framework for Network Intrusion Detection in Cloud," 2016 IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, pp. 512-516.

[16]. David Chappell, (2015). "Introducing Azure Machine Learning", A GUIDE FOR TE-ComHNICAL PROFESSIONALS, Sponsored by Microsoft Corporation, 2015.

[17]. https://studio.azureml.net/