# "Intrusion Detection System Evaluation "A Comparative Study of Machine Learning Algorithms

**Praveen Kumar Shukla[1] and Dr C S Raghuvanshi[2] and Dr Hari Om Sharan[3]**

[1]Department of electronic & communication Engineering, Rama University, Kanpur 209217, INDIA
**Email Id:** praveenshukla26@gmail.com
[2] Department of Electronic & communication Engineering, Rama University, Kanpur 209217, INDIA
**Email Id:** drcsraghuvanshi@gmail.com
[3] Department of Electronic & communication Engineering, Rama University, Kanpur 209217, INDIA
**Email Id:** deanengineering@ramauniversity.ac.in

**Abstract**: The need for cheaper and faster delivery in the electronics industry has increased as a result of information technology advancements. The quick development of technology not only makes life simpler but also raises several security concerns. The number of attacks conducted online has increased as the Internet has developed through time. One of the supporting layers that can be used for information security is the intrusion detection system (IDS). IDS offers a clean atmosphere for conducting business and steers clear of shady network activity. The security on the user's end of web transactions is the most difficult task in the construction of an e-commerce system. This study examined intrusion detection security techniques. Continuous monitoring of intrusion detection is required for further technological adaptation, and as a result, presents a comparative comparison of adaptive artificial intelligence-based intrusion detection algorithms. This work shows how reinforcement learning (RL) and regression learning-based intrusion detection systems (IDS) can be used to solve extremely difficult issues, such as choosing input features and taking limited resources into account.

**Keywords**: reinforcement learning, linear regression, logistic regression, intrusion detection, KDD dataset.

## 1. INTRODUCTION

Electronic commerce (E-commerce) is the only way to meet the needs of business organizations in the modern, digital age. E-commerce accelerates delivery by offering high-quality goods at lower prices. It is also known as paperless business information sharing at times. It is well-liked by business partners due to its multifunctional characteristics, such as non-cash payments, 24-hour services, better marketing management, pre-sales and after-sales services, and quick and dependable client communication. It serves as a unified platform for information sharing, strategy upkeep, and support of global business operations. Aside from these benefits, there are two key technological and non-technical shortcomings with e-commerce. Due to ineffective algorithm implementation, the security of the system is its primary technical drawback. It is quite challenging to offer security

for web transactions at the user end. The owner and user will suffer an unsupportable financial loss. The reason that causes this loss in e-commerce is the downtime brought on by an intrusion attack. As a result, this study uses intrusion detection algorithms to analyse security algorithms.

## 2. LITERATURE REVIEW

| PAPER TITLE | RESEARCH TECHNIQUE | FUTURE SCOPE |
|---|---|---|
| **Semi-supervised multi-layered clustering model for intrusion detection. Digital Communications and Networks 4(4): 277-286.** | In this paper SMLC (an ensemble model of multiple randomized layers using K- means algorithm) is made and then this model result is compared | SMLC has a relatively high testing time. In future work should be done to improve the scalability of SMLC, automation of its parameter |

_____

| | with the known ML algorithms such as random forest, adaboost, bagging to detect the accuracy of the model. The SMLC model reduces the computational cost and can process large amounts of data. | tuning process and reduction of its testing time. |
|---|---|---|
| **Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection-Procedia Computer Science 89(1): 117-123.** | In this paper four ML algorithms are used on a NSL-KDDdataset and their accuracy, F-scores, precision, recalls are compared and out of the four models Random forest model has given the best results. | In future work can be extended by considering the classifiers for multiclass classification and considering only the important attributes for the intrusion detection |
| **IDS for IoT-based smart environments: a survey Journal of Cloud Computing 7 (1): 21** | In this paper the IDS system is integrated with the IOT smart applications to prevent attacks and security issues. Here the PCA (Principal component analysis) algorithm is used in IDS as it is a lightweight algorithm for IOT-based environment and descriptive multivariate method for handling quantitative data and can be extended to deal with mixed measurement level data. | In future work can be done in the areas to reduce the monitored traffic and increase the processing time and automate the detection process without human intervention. |
| **Intrusion detection in smart cities using** | In this paper two main methods are used. One is the | In future work can be done to reduce the error percentage |

| **Restricted Boltzmann Machines." Journal of Network and Computer Applications 135 (1): 76-83.** | RBM(Restricted Boltzmann model) for feature extraction and the other is classifier model for attack detection. | and detect variety of attacks. |
|---|---|---|

## 3. MATERIALS AND METHODOLOGY

### Step1: Import libraries and load dataset



### Step2: Data pre-processing

**Scaling numerical Attributes**: Taking away the mean and scaling to one variance will normalize the characteristics. Statistics regarding the cases in the training phase are independently targeted and adjusted by computing each attribute. The data that follows are then modified using the standard error and average that were created earlier. Many approximations require dataset uniformity; if the specific qualities do not match standard data that is normally distributed, they may perform poorly.

Pre-processing of Data

- Dataset had **45 attributes** and **175341 rows**.

- The dataset has 81173 rows and 45 attributes after the null values were removed.

- Utilizing given data type information from the features dataset, the data type of attributes is converted.

**1303**

_____





## Step3: Data cleansing, normalization, and null value detection

In Python Label Encoding, the category value must be changed to a numeric value in the range of 0 to the class labels' total number, minus one. if the outcome of the category variable.

### *Data Normalization*

- Normalization of Data 58 Data Frame's numeric columns are scaled from 0 to 1 using the Min Max Scaler.
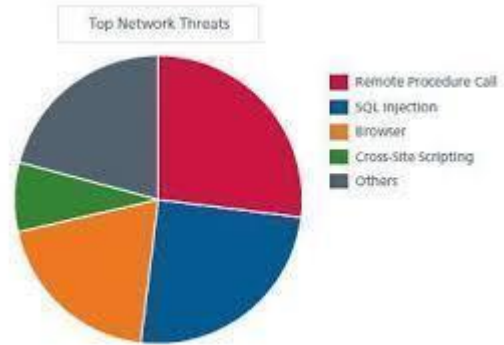
### *Binary Classification Preparation*

- The Data Frame is duplicated for the Binary Classification process.

- *The 'label' property is divided into two groups: 'normal' and 'abnormal'.*

- *Using Label Encoder (), the column 'label' is encoded, and the resulting encoded labels (0,1) are preserved in the column itself.*

- *61 columns and 81173 rows in a binary dataset*

### *Multi-class Classification Preparation*

- The Data Frame is duplicated for the Multi-class Classification process.

- *Nine categories are listed under the 'attack_cat' attribute: 'Analysis', 'Backdoor', 'DoS', 'Exploits', 'Fuzzers', 'Generic', 'Normal', 'Reconnaissance', and 'Worms'.*

- *Label Encoder () is used to encrypt attack_cat, and the accompanying encoded labels (0,1,2,3,4,5,6,7,8) are kept in the label attribute.*

- *attack_cat is encoded using one-hot.*

- Multi-class Dataset — 81173 rows, 69 columns



## Step4: Classify the type of attacks

_____

**Step5: Feature Extraction**:

This is the procedure of reducing the number of variables in a statistical model. The number of variables should be kept to a minimum to decrease overall computation costs and, in some cases, to improve model performance. The input parameters with the strongest correlation to the target attribute are selected after statistical analysis is used to assess the relationship between each input parameter and the goal parameter. Although the statistical measures utilized are influenced by the type of data in both the dependent and independent variables, these processes can be simple and efficient. The random forest gets estimates from each decision tree and calculates the right answer based on the number of votes cast, as opposed to relying just on one decision tree. Compared to other techniques, training takes less time with this one. Even with a large dataset, it runs quickly and accurately predicts the outcome. Even in the absence of a sizable volume of data, accuracy can still be maintained.

***Selection of features***

- 'bin_data' has 61 characteristics total.

- 'multi_data' has 69 properties total.

- For feature extraction, the Pearson Correlation Coefficient approach is employed.

- The attributes with a correlation coefficient of more than 0.3 to the target attribute label were chosen.

- The attributes with a correlation coefficient of more than 0.3 to the target attribute label were chosen

- Number of 'bin_data' attributes after feature selection: 15

- "rate,""sttl,""sload,""dload,""ct_srv_src,""ct_state_ttl,""ct_dst_ltm,""ct_dst_src_ltm,""ct_src_ltm, ""ct_srv_dst," "state_CON," "state_INT," and "label."

- After feature selection, the number of "multi_data" attributes is 16

- "dttl," "swin," "dwin," tcprtt," synack," ackdat," label," proto_tcp,"proto_udp,"service_dns," state_CON,"state_FIN,"attack_cat_Analysis,"attack_cat_DoS,"attack_cat_Exploits,"andattack_cat_Normal."



**Step 6: Splitting the Dataset into the Train & Test sets**:

we separate our sample into a test set and a train set. This important data pre-processing step increases our classification model's efficacy. Suppose we train our classification model on a single sample and validate it on a second sample. Consequently, our model will struggle to comprehend how the two models relate to one another. The model's performance will deteriorate if we give it a new dataset after it has received enough training and has a high training accuracy. A substantial amount of effort has been devoted to the development of a classification model that exhibits strong performance on the test & training datasets alike. The train set was included in the sample used to develop the classification model, and the result is now known for these samples. The term "test set" refers to a subset of the sample that is used to forecast output when the model is tested.

**Splitting Dataset into Training and Testing**

- Randomly splitting the **bin_data** in **80% for training** and **20% for testing**.

- Randomly splitting the **multi_data** in **70% for training** and **30% for testing**.
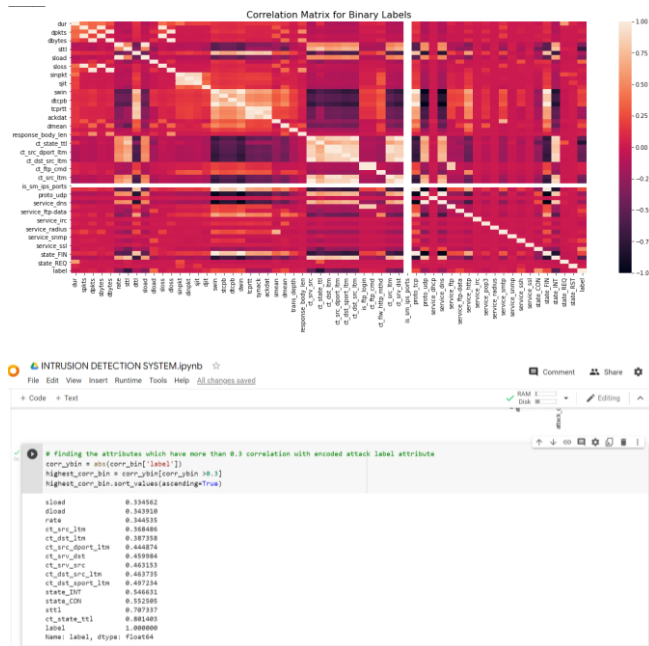
- Target feature — *label*

**Step7: Comparison of Linear Regression, Logistic Regression and RL.**

**Linear regression models**

It is possible to look at the link between one or more independent variables and a continuous dependent variable. If there is only one independent variable and one dependent variable, then simple linear regression is used. When there are more than two factors, though, multiple linear regression is used instead of simple linear regression. In all types of linear regression, the least squares method is used to find the line that fits a set of data the best. Logistic regression, like linear regression, can be used to foretell numbers for categorical variables. This method is used to evaluate the correlation between several predictors and a single outcome. Categorical variables can take on the values true, false, yes, no, 1, 0, and other values. The S-curve is transformed into a straight line by the logit function, and the unit of measurement in linear regression is a probability. In regression analysis, both models are used to make predictions about the future. However, linear regression is often easier to understand. In addition, whereas a smaller sample size is sufficient for linear regression to accurately represent values across all answer categories, a larger sample size is necessary for logistic regression. Should a more extensive, representative sample be lacking, the model's statistical potential to identify a noteworthy effect may be compromised.

**Linear Regression Model**

*Binary Classification*

- Accuracy — **97.80720665229443**

- Mean Absolute Error — **0.021927933477055742**

- Mean Squared Error — **0.021927933477055742**

- Root Mean Squared Error — **0.1480808342664767**

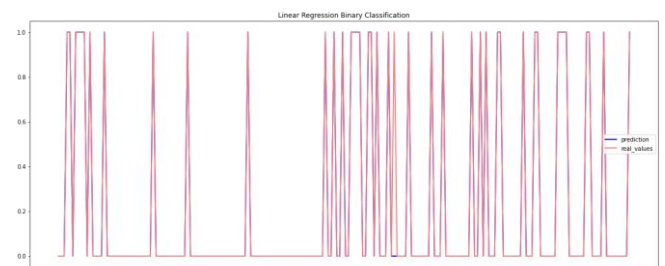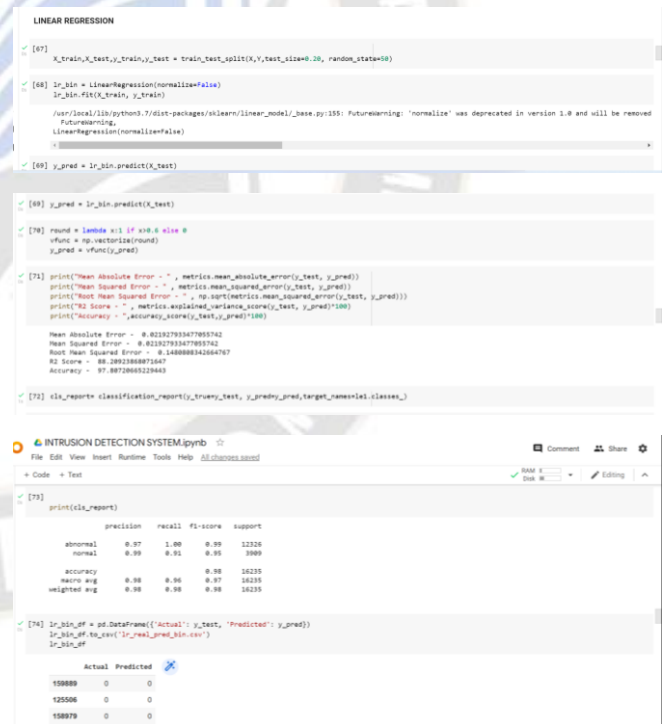- R2 Score — **88.20923868071647**

*Multi-class Classification*

- Accuracy — **95.12976346911958**

- Mean Absolute Error — **0.06824901445466491**

- Mean Squared Error — **0.12146846254927726**

- Root Mean Squared Error — **0.3485232596962178**

- R2 Score — **91.82055676180129**

_____

### Logistic regression

Among the family of supervised machine learning models is logistic regression. It is also considered a discriminative model, meaning that it aims to differentiate between multiple classes. Naive Bayes and other generative algorithms can make examples of the class they are trying to predict, but this one can't (for example, a cat picture). We already talked about how the model's beta values are made by logistic regression, which maximises the log likelihood function. This slightly alters when looking at it from a machine learning standpoint. In machine learning, the global maximum is found using gradient descent, with the loss function being the negative log likelihood. This is merely a different method to get the same approximations as before. Moreover, logistic regression may lead to overfitting, especially if the model has a large number of predictor variables. In high dimensionality models, regularisation is usually employed to penalise parameters with large coefficients. For additional information on the logistic regression machine learning model, go visit Scikit-learn (link is external to IBM).

### Logistic Regression Model

#### Binary Classification

- Accuracy — **97.80104712041884**

- Mean Absolute Error — **0.0219952879581152**

- Mean Squared Error — **0.0219952879581152**

- Root Mean Squared Error — **0.1482886671186019**

- R2 Score — **88.17947258428785**

#### Multi-class Classification

- Accuracy — **97.58952036793693**

- Mean Absolute Error — **0.060077201051248356**

- Mean Squared Error — **0.18056011826544022**

- Root Mean Squared Error — **0.42492366169165047**

- R2 Score — **87.87674567880146**



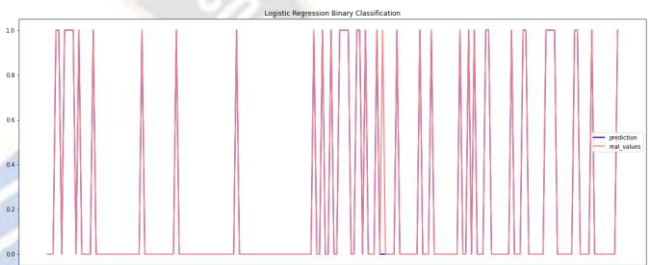Logistic regression is commonly applied to problems involving classification and forecasting through the use of cases. These use cases consist of:

**•Fraud detection:**

Teams can use logistic regression models to find data abnormalities that point to fraudulent activity. Banking and other financial organisations may discover that specific behaviours or qualities are more often linked to fraudulent operations in an effort to better protect their consumers. SaaS-based organisations also employ these techniques to eliminate fictitious user accounts from datasets during business performance data analysis.

**•Disease prediction:**

This analytics technique can be implemented in medicine to predict the likelihood that a given population will become unwell or sick. Health care facilities can arrange for prophylactic care for individuals who are more susceptible to a specific disease.

**•Churn prediction:**

Certain behaviours may signal churn in different organisational activities. If high achievers are in danger of leaving the organisation, human resources and management teams may be interested in learning about it. Information of this kind could start conversations regarding

**1307**

_____

___

the company's compensation policies or culture. Alternatively, the sales staff would like to know which of its clients might choose to work with another company. Teams might be inspired by this to develop a retention plan in order to prevent revenue loss.

Types of logistic regression Logistic regression models based on categorical responses come in three varieties.

**•Binary logistic regression:**

This method's dependent variable, or answer, is dichotomous, meaning that there are only two possible outcomes (e.g., 0 or 1). It can determine the malignancy of a tumour or whether an email is spam, among other applications. This technique is widely used in the field of logistic regression and is widely recognised as a top-tier classifier for binary data.

**•Multinomial logistic regression:**

Within this particular variant of logistic regression, the dependent variable is open to a maximum of three values, none of which are sequentially arranged. To improve their ability to sell their films, for instance, film studios endeavour to predict the genre of film that a given audience is most likely to enjoy. To determine the degree to which age, relationship status, gender, and age influence the genre of film that an individual prefers, the studio may implement a multinomial logistic regression model. As a result, the studio may opt to focus its marketing endeavours on the specific demographic that is relevant to a given film.

**•Ordinal logistic regression:**

Despite the set order of values shown, this type of logistic regression model is used when the response variable has three or more possible outcomes. Ordinal responses can be expressed as grading schemes or rating scales ranging from 1 to 5.
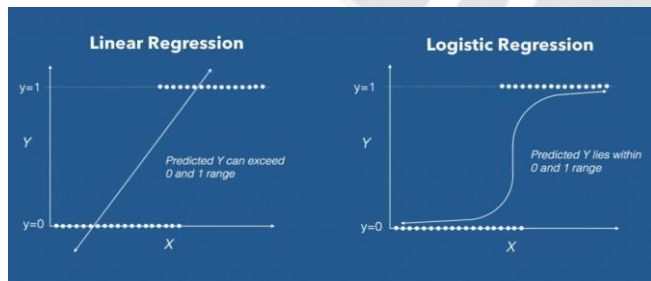


*Multi-class Classification*

- Accuracy — **97.59362680683311**
- Mean Absolute Error — **0.059912943495400786**

- Mean Squared Error — **0.17941031537450722**
- Root Mean Squared Error — **0.42356854861345317**
- R2 Score — **87.93449282205455**

- **SVM algorithm**

1. **Input:** The algorithm takes a labelled dataset as input, where each data point consists of a set of features (X) & a corresponding binary class label (Y) indicating one of two classes (e.g., 0 or 1).

2. **Data Preparation:** If needed, feature scaling and data normalization can be performed to ensure that all features have a similar scale, which can improve the convergence and performance of the algorithm.

3. **Feature Space Transformation Kernel Trick - Optional:** The initial feature space can be changed into a higher-dimensional space using a kernel function in SVM. This enables SVM to identify a decision boundary that may have been nonlinear in the original space but is now linear in the modified space. The linear kernel, polynomial kernel, sigmoid kernel, and radial basis function RBF kernel are examples of common kernel functions.

4. **Margin Maximization:** Finding the hyperplane with the largest margin between the two classes is the primary goal of SVM. The margin is the distance between the hyperplane and the closest data points of each class support vectors. The optimal hyperplane is the one that achieves the largest margin while correctly classifying the data points.

5. **Soft Margin Optional**: SVM introduces slack variables to allow for a soft margin in situations where the data is not perfectly separable. These slack variables allow some data points to be misclassified or fall within the margin region, but with a penalty for misclassification. The objective is to balance the margin maximization and the classification error.

6. **Optimization Problem:** The SVM technique frames the issue as a convex optimisation problem in order to determine the ideal weights and biases for the hyperplane. The optimisation seeks to maximise the margin and minimise the classification error of incorrectly categorised data points.

7. **Kernel Parameters Optional:** In the event that a kernel function is employed, the SVM algorithm may further entail fine-tuning the kernel parameters, such as the polynomial kernel's degree or the RBF kernel's

**1308**

width, in order to maximise performance on the training set.

8. **Training:** In the training phase, the algorithm finds the ideal hyperplane parameters by solving the optimisation issue using methods such as quadratic programming or Sequential Minimal Optimisation (SMO).

9. **Prediction:** The model can be used to predict the class label for fresh data points once it has been trained. The predicted class label is determined by comparing the signed distance of the data point to the hyperplane. If the signed distance is positive, the data point belongs to one class; otherwise, it belongs to the other class.

10. **Output:** For binary classification tasks, the Support Vector Machine technique provides an output of the learnt model parameters (weights and bias).It is well known that SVM can handle high-dimensional data and find difficult decision boundaries, making it a reliable classification system. When the data are clear and the categories are distinct, this method shines. Using techniques such as One-vs-One (OvO) or One-vs-Rest OvR strategies, SVM can be adapted to perform multiclass classification. In addition, the Support Vector Regression SVR variant of SVM can be used for regression assignments.

## 4. RESULTS AND DISCUSSION

The assessment of the adaptive neural network approach to intrusion detection involved conducting three experiments using the prototype application. The effectiveness of several algorithms, such as SVM, linear regression, & logistic regression, was assessed in these trials.

The results clearly demonstrated that the support vector machine consistently provided higher accuracy compared to both linear and logistic regression models. This suggests that the adaptive neural network approach, when implemented using a support vector machine, holds significant promise for improving IDS ' effectiveness and accuracy.

## 5. CONCLUSION

IDS are essential for protecting computer and network-based systems because they continuously monitor protocol traffic for possible breaches and assaults. The adaptability of IDS allows for adjustments in response to both external and internal threats. In this study, we conducted a comparative analysis of IDS using linear logistic regression & SVM.

In conclusion, the experiments conducted using the prototype application to assess the viability of the adaptive neural network approach for intrusion detection revealed valuable insights. Specifically, the support vector machine algorithm consistently outperformed linear and logistic regression models in terms of accuracy. This finding underscores the potential effectiveness of the adaptive neural network approach for enhancing IDS.

These findings imply that more investigation and advancement in this area may result in IDS that are more reliable and accurate. Implementing adaptive neural networks, particularly with SVM, holds promise for addressing the evolving challenges in cybersecurity and bolstering our ability to detect and respond to intrusions effectively. However, additional investigations and real-world testing are warranted to fully harness the capabilities of this approach and validate its practicality in various security contexts.

## ACKNOWLEDGMENT

## REFERENCES

[1] Aburomman, A. A., &Reaz, M. B. I. (2016) "Ensemble of binary SVM classifiers based on PCA and LDA feature extraction for intrusion detection."Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC): 636-640.

[2] Al-Jarrah, O. Y., Al-Hammdi, Y., Yoo, P. D., Muhaidat, S., & Al-Qutayri, M. (2018) "Semi-supervised multi-layered clustering model for intrusion detection." Digital Communications and Networks 4(4): 277-286.

[3] Al-Yaseen, W. L., Othman, Z. A., &Nazri, M. Z. A. (2017) "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system." Expert Systems with Applications 67(1): 296-303.

[4] An, X., Su, J., Lü, X., & Lin, F. (2018) "Hypergraph clustering model-based association analysis of DDOS attacks in fog computing intrusion detection system." EURASIP Journal on Wireless Communications and Networking 249 (1): 1-9.

[5] Belavagi, M. C., &Muniyal, B. (2016) "Performance evaluation of supervised machine learning algorithms for intrusion detection." Procedia Computer Science 89(1): 117-123.

[6] Elrawy, M. F., Awad, A. I., &Hamed, H. F. (2018) "Intrusion detection systems for IoT-based smart

_____

environments: a survey." Journal of Cloud Computing 7 (1): 21

[7] Elsaeidy, A., Munasinghe, K. S., Sharma, D., &Jamalipour, A. (2019) "Intrusion detection in smart cities using Restricted Boltzmann Machines." Journal of Network and Computer Applications 135 (1): 76-83.

[8] Ganapathy, S., Kulothungan, K., Muthurajkumar, S., Vijayalakshmi, M., Yogesh, P., &Kannan, A. (2013) "Intelligent feature selection and classification techniques for intrusion detection in networks: a survey." EURASIP Journal on Wireless Communications and Networking (1): 271.

[9] Gautam, R. K. S., &Doegar, E. A. (2018) "An Ensemble Approach for Intrusion Detection System Using Machine Learning Algorithms." International Conference on Cloud Computing, Data Science & Engineering (Confluence): 14-15.

[10] Ghasempour, A. (2019). Internet of Things in Smart Grid: Architecture, Applications, Services, Key Technologies, and Challenges. Inventions 4(1): 22.