JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

COMPARATIVE ANALYSIS OF MALWARE DETECTION DATASETS USING DIFFERENT MACHINE LEARNING CLASSIFIERS

Result Analysis for Accuracy in ML Classifiers Techniques

[1]Mr. Swapnil S. Chaudhari, [2]Dr. Satish N. Gujar, [3]Dr. Farhat Jummani

[1]Research scholar, [2]Research Guide, [3]Research Guide

[1]Computer Engineering Department,

Abstract : The region of Malware Detection is not new, Malware is a malicious code that is done to harm a computer or network. There are three main methods are in the market to sense malware: Signature-based, Behavioral based, and Heuristic ones methods, but to find the more accurate result in anomaly-based Malware detection Machine Learning algorithms are used. Therefore this study is based on choosing the best machine learning classification methods for the detection of malware. In this work, the implementation of such an efficient and versatile MDS is a profound learning-oriented manner. In This comparative Analysis of Malware Detection We took two different datasets of Malware detection online as dataset_malwares.csv, Data.csv of Rows:19243, Columns:79 & Rows:10539, Columns:57 also applied four different Machine Learning classification techniques as naive Bayes classification, the random forest classification, Decision Tree Classifier & Linear SVC Classifier for to find better accuracy. Through both, the test conducted on Classifiers Machine Learning is verified to be successful for MDS. The best classification method is Random forest classification which has classified the subjects with an average of 97.955% accuracy. This can prove to be a useful screening classification tool for detecting malware in systems (MDS) and networks (NMDS).

***Index Terms/Keywords* - Accuracy in Classifiers, Malware Analysis, Malware Classification, Malware Detection, ML Classifiers, Random Forest Classifier**

I. INTRODUCTION

In today's computer environment, data and application protection is most important due to the development and global enhancement in IT. The sharing of communication and information technologies that generate new value-added services through various cyber threats, and Malwares. They have built online services for the benefit. Nonetheless, cyber security risks also are growing because as Internet contact points are increasing. The Malware Detection system (MDS) is an important safety concern today. A Network Malware Detection System (NMDS) assists server administrators to identify network security vulnerabilities within their operations. However, when a stable and powerful NMDS is designed for unpredictable and unforeseeable attacks, several problems occur. For big organizations, device security is an integral part of now days. Frameworks for Intrusion Malware Detection system (MDS) are becoming exceptional for positive guarantee against constantly changing assaults in terms of size and complexity. They must be easy and reliable to control with information integrity, insulation and accessibility and low maintenance costs. Every time, different changes are connected to MDS to learn and cope with new assaults. Arwa Aldweesh et. al. [1] the main simulations evaluating deep learning for malware detection were evaluated and compared, as well as the current sample was based on historical ones. It provided an interesting fine-grained categorization that considered different modelling aspects, namely data input, recognition, implementation, and strategies for assessment. This thus offered an in-depth analysis of the relevant scientific investigations in auditory learning style IDS. R. Vinayakumar et. al. [26] describes the deep Neural Network (DNN), a form of deep learning model, to create scalable and efficient MDS for the identification and classification of unexpected and unpredictable cyber-attacks. The constant change in network behavior system's and the rapid evolution of attacks make it important to analyze various datasets that are created by static and dynamic approaches over the years. This form of research helps determine the best algorithms that can work efficiently to detect potential cyber-attacks. A thorough assessment of DNN applications and some other classical machine-learning classification algorithms is seen on various publicly available comparison malware datasets. There are different type of Machine Learning concepts and formats. One of them is classification technique. Basically, classification is about identifying in which set of categories a certain observation belongs in the system datasets. Classifications are normally belonging to supervised learning techniques in the field of Machine Learning. A typical classification is Spam detection in e-mails in Gateway – the two possible classifications in this case are either "spam" or "no spam". The two most common classification algorithms are the naive bayes classification, the random forest classification, Decision Tree Classifier & Linear SVC Classifier In this study we have concentrated on these four classification techniques and took observations on two different datasets from Malware Detection.

The Decision Tree classifier

The basic classifier technique is the Decision tree classifier. Decision Trees be a type of Supervised Machine Learning. It basically builds classification models and rules in the form of a tree structure. The dataset is broken down and paired like trees into smaller subsets and gets detailed by each leaf. where the data is simultaneously split according to a definite parameter. The tree can be clarify by two entities, namely decision tree nodes and leaves. The leaves are the decisions or the ending outcomes. And the choice nodes are where the data is split. It would be compared to a study, where each question has an effect on the next questions. Let's presume the following case: If your age is greater tan 18 then you go for the Vaccination like this. Basically, by going from one leave to another, you get closer to your result. Classification trees are of (Yes/No types) & Regression trees are of (Continuous data types). Regression trees pass on to an algorithm where the goal variable is and the algorithm is used to forecast its value. As an exemplar of a regression type problem, you may want to predict the advertising prices of a residential house, which is a continuous dependent variable. This will depend on both continuous factors like square footage as well as categorical features like the style of home, area in which the property is located and so on.

When to use Classification and Regression Trees models: Classification trees are used when the dataset wants to be split into classes which belong to the reaction variable. In many cases, the classes Yes or No (1,0). In other words, they are now two and mutually exclusive. In some cases, there may be other than two classes in which case a alternative of the classification tree algorithm is used. Regression trees, on the other tender, are used when the response variable is uninterrupted. For instance, if the response variable is impressive like the price of a property or the temperature of the day, a regression tree is used. In other words, regression trees are used for prediction-type troubles while classification trees are used for classification-type problems.

The Random Forest classification

Random forest is a quietly good classifier, often time used and also often very efficient in correctness. It is an ensemble classifier ended using many decision tree models in system. There are ensemble models that merge the different results. The random forest model can together run regression and classification models. Basically, it split the data set into subsets and after that runs on the data. Random forest models scuttle efficient on large datasets, since all calculative result can be split and thus it is easier to run the model in equivalent. It can handle thousands of input variables without variable cutting. It computes proximities flanked by pairs of cases that can be used in clustering, locating outliers or (by scaling) give interesting views of the data. A random forest is an ensemble learning method where multiple decision trees are constructed and then they are compound to get a more accurate prediction. If there is one method in ML that has grown in popularity over the last few years, then it is the idea of random forests. The concept has been around for longer than that, with several different people inventing variations

The Naive Bayes classifier

The Naive Bayes classifier is based on previous knowledge of conditions that might recount to an event. It is based on the Bayes Theorem. There is a physically powerful independence between features assumed. It uses uncompromising data to calculate ratios between measures. The benefits of Naive Bayes are different. It can straightforwardly and fast predict classes of data sets. Also, it can predict multiple classes. Naive Bayes performs superior compared to models such as logistic regression and there is a lot less training data needed.

Support Vector Machines

Support Vector Machine or SVM is one of the nearly all popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, mainly, it is used for Classification problems in Machine Learning. The ambition of the SVM algorithm is to create the finest line or decision boundary that can separate out n-dimensional space into classes so that we can straightforwardly put the new data point in the acceptable category in the future. This best decision border line is called a hyper plane. SVM prefer the extreme points/vectors that help in creating the hyper plane. These tremendous cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Linear SVM: Linear SVM is used for linearly separable data, which funds if a dataset can be classified into two program by using a single straight line, then such data is period as linearly separable data, and classifier is used called as Linear SVM classifier.

Non-linear SVM*:* Non-Linear SVM is used for non-linearly estranged data, which means if a dataset cannot be classified by using a without delay line, then such data is period as non-linear data and classifier used is called as Non-linear SVM classifier.

1.1 Facts with Supervised Learning:

- Supervised learning algorithms are trained using labeled data using datasets
- Supervised learning model takes direct feedback to check if it is predicting correct output or not

- Supervised learning model predicts the output based on ML Algorithms.
- In supervised learning, input data is provided to the model along with the output based on ML Algorithms.
- The goal of supervised learning is to train the model so that it can predict the output when it is given new data.
- Supervised learning needs supervision to train the model .
- Supervised learning can be categorized in Classification and Regression models.
- Supervised learning can be used for those cases where we know the input as well as corresponding outputs.
- Supervised learning model produces an accurate result for predictions in systems.
- Supervised learning is not close to factual Artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output.
- It includes diverse algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc.

II. Objective

The aim of the study is to develop a different Machine Learning classification model for classifying Legitimate and malware labels. Various classification models are used and compared based on classification accuracy for Naive Bayes classification, the Random Forest classification, Decision Tree Classifier & Linear SVC Classifier.

III. Datasets

The dataset for this study we uses open source datasets. In This comparative Analysis of Malware Detection We took two different datasets of Malware detection online as dataset_malwares.csv, Data.csv of Rows:19243, Columns:79 & Rows:10539, Columns:57 also applied four different Machine Learning classification techniques as Naive Bayes classification, the Random Forest classification, Decision Tree Classifier & Linear SVC Classifier for to find better accuracy.

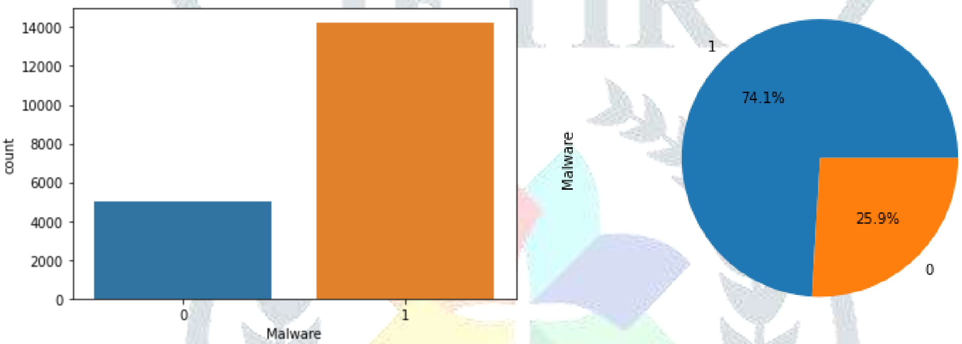

Fig. 1. Malware & Legitimate Count in dataset_malwares.csv

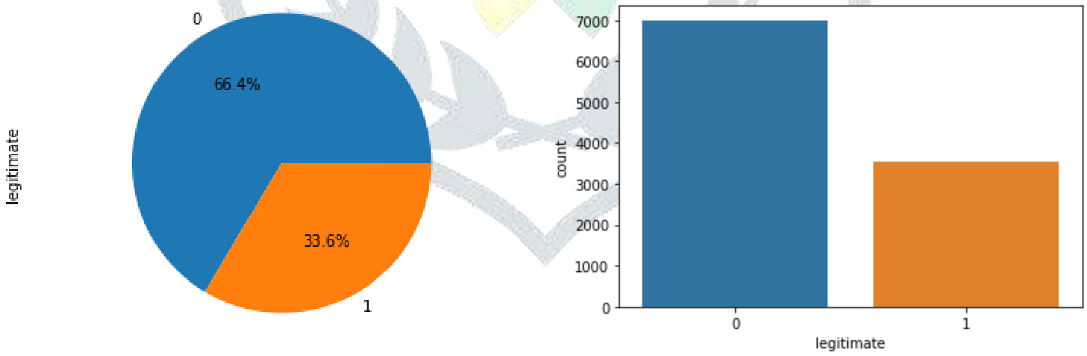

Fig. 2. Malware & Legitimate Count in data.csv

IV. Analysis

The Analysis is headed by data cleaning and processing first for both the datasets. ML classification models have been used as a case in this article to explain the process better of system. Analysis done on four different Machine Learning classification techniques as Naive Bayes classification, the Random Forest classification, Decision Tree Classifier & Linear SVC Classifier. Correlation is an indication about the changes between two variables herewith we have introduces following correlations for both the datasets.

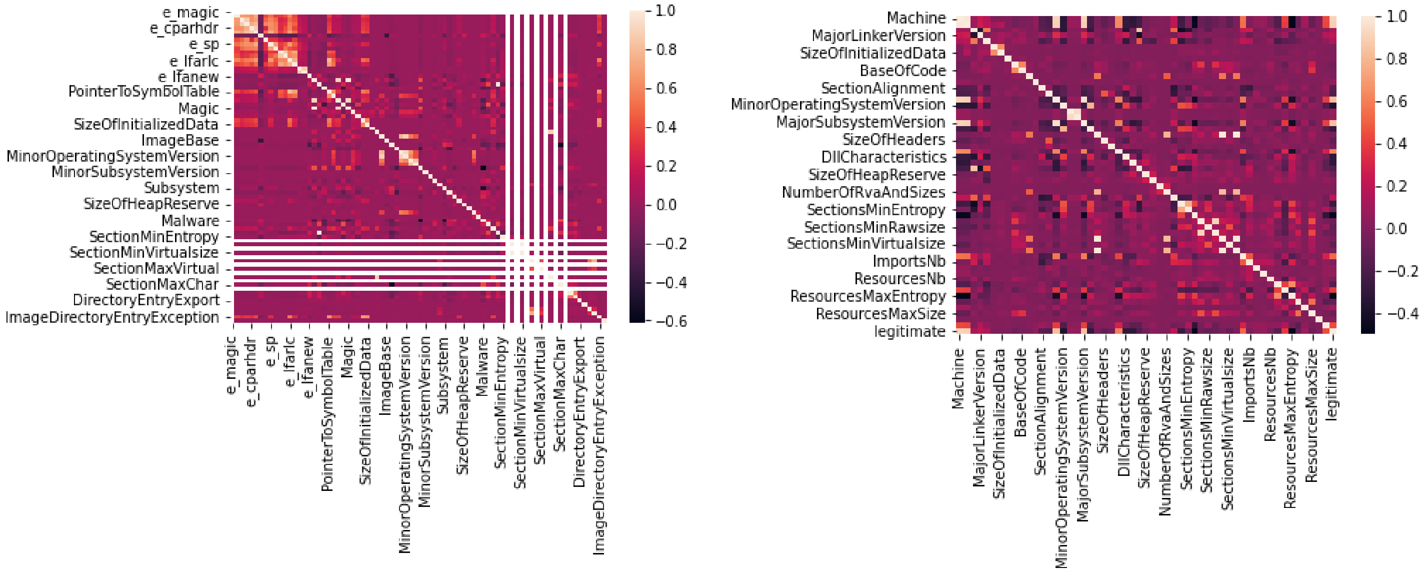

Fig. 3. Malware & Legitimate correlations with labels in Both Datasets

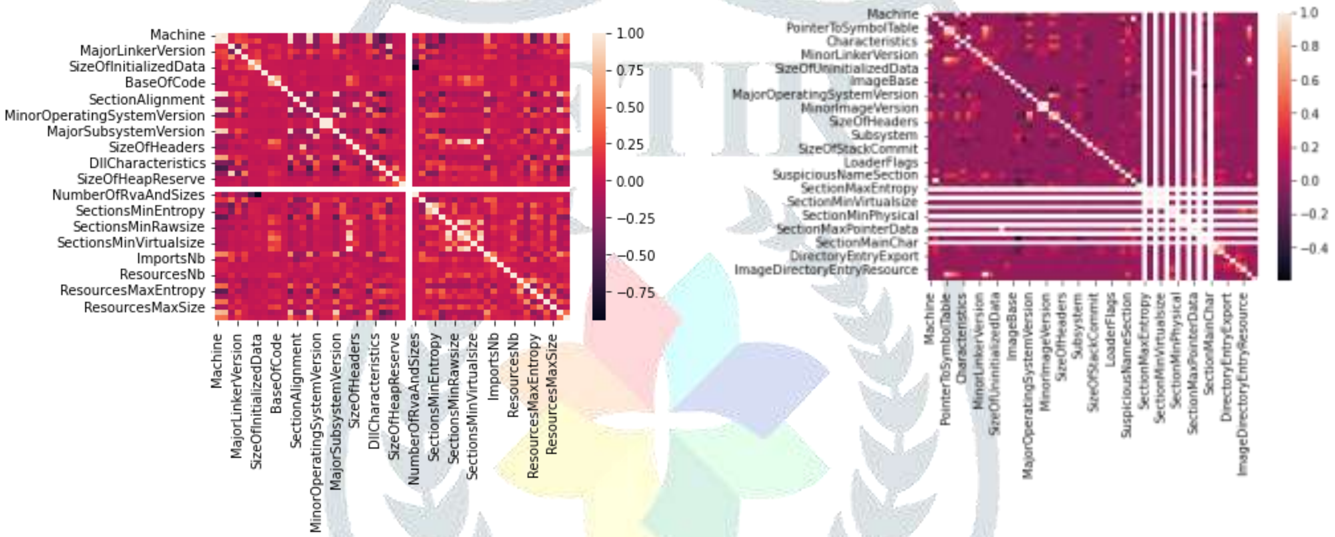

Fig. 4. Malware & Legitimate Test Data correlations with labels in Both Datasets

From the above output of correlation matrix, we can see that it is symmetrical i.e. the bottom left is same as the top right. It is also observed that each variable is positively correlated with each other.

V. RESULTS AND DISCUSSION

Results are shown in the Table 1

TABLE I: ML Four Classification Methods Accuracy results with Both Datasets

Dataset_Name	GaussianNB_ Acuracy	RandomForest_ Acuracy	DecisionTree_ Acuracy	SupportVector_ Acuracy
dataset_ malwares.csv	32.0083	98.28527	40.32216	96.092085
Data.csv	36.1006	97.62808	40.08539	95.986336
Average	**34.0544**	**97.95668**	**40.20378**	**96.039211**

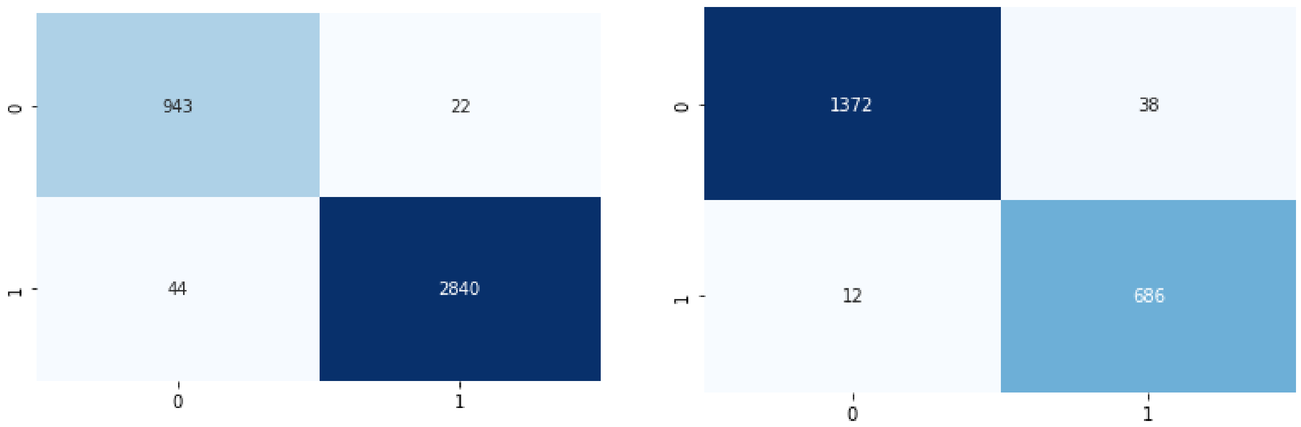

Fig. 5. Heat Maps of Confusion Matrix for both Datasets

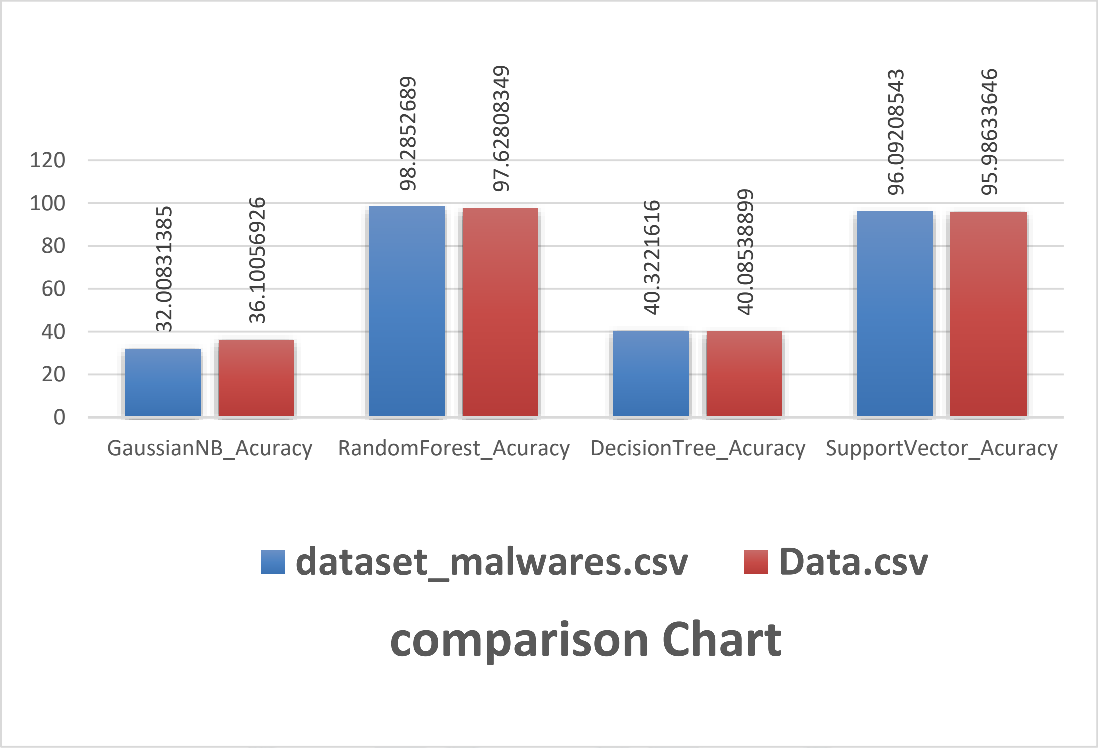

Fig. 6. Comparison Chart final results

VI. Conclusion & Discussion

With the available data set, the best classification methods are Random Forest and Support Vector Machine which have classified the subjects with Random Forest (Average) 97.95668% accuracy & with Support Vector Machine (Average) 96.039211% Accuracy. This can prove to be a useful screening & Classification tool for detecting malware in systems and Networks. This can be used as an effective screening tool and further mechanisms like signature detection and anomaly based detection can be deployed to improve the Malware detection system and to improve Malware Detection in Networks also.

Acknowledgment

I express our sincere thanks to Dr. Farhat Jummani for her kind co-operation and Valuable Guidance. I also express our sincere thanks to Dr. Satish N. Gujar for her kind Cooperation in research work

References

[1] Aldweesh Arwa, Abdelouahid Derhab, and Ahmed Z. Emam. (2019) "Deep Learning Approaches For Anomaly-Based Intrusion Detection Systems: A Survey, Taxonomy, And Open Issues." Knowledge-Based Systems 189 (2019): 105124.

[2] Alrawashdeh, Khaled, and Carla Purdy. (2016) "Toward an Online Anomaly Intrusion Detection System Based on Deep Learning," 2016 15th IEEE international conference on machine learning and applications (ICMLA). IEEE, 2016.

[3] Althubiti, Sara, et al. (2018) "Applying Long Short-Term Memory Recurrent Neural Network for Intrusion Detection." SoutheastCon 2018. IEEE, 2018.

[4] Anish Halimaa A,Dr. K.Sundarakantham, 2019, MACHINE LEARNING BASED INTRUSION DETECTION SYSTEM ,Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8, PP 916-920"

[5] Anna L. Buczak, Member, IEEE, and Erhan Guven,,2016,A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection, IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 18, NO. 2, SECOND QUARTER 2016, PP 1153-1176

[6] Cordero, Carlos Garcia, et al. (2016) "Analyzing Flow-Based Anomaly Intrusion Detection Using Replicator Neural Networks." 2016 14th Annual Conference on Privacy, Security and Trust (PST). IEEE, 2016

[7] Dong Bo, and Xue Wang. (2016) "Comparison Deep Learning Method to Traditional Methods Using for Network Intrusion Detection." 2016 8th IEEE International Conference on Communication Software and Networks (ICCSN). IEEE, 2016.

[8] Farahnakian Fahimeh, and Jukka Heikkonen. (2018) "A Deep Auto-Encoder Based Approach for Intrusion Detection System." 2018 20th International Conference on Advanced Communication Technology (ICACT). IEEE, 2018.

[9] Hassan Azwar,Muhammad Murtaz,Mehwish Siddique,2018,Intrusion Detection in secure network for Cybersecurity systems using Machine Learning and Data Mining,2018 IEEE 5th International Conference on Engineering Technologies & Applied Sciences, 22- 23 Nov 2018, Bangkok Thailand

[10] Hemant Dhamija and Ajay K. Dhamija (2021) "Malware Detection using Machine Learning Classification Algorithms" International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 17, Number 1 (2021), pp. 1-7, © Research India Publications ,http://www.ripublication.com

[11] Imamverdiyev, Yadigar, and Fargana Abdullayeva. (2018) "Deep Learning Method for Denial of Service Attack Detection Based On Restricted Boltzmann Machine." Big Data 6.2 (2018): 159-169.

[12] Ishaque, Mohammed, and Ladislav Hudec. (2019) "Feature Extraction Using Deep Learning for Intrusion Detection System." 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS). IEEE, 2019.

[13] Javaid, Ahmad, et al. (2016) "A Deep Learning Approach for Network Intrusion Detection System," Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS). 2016.

[14] Kazi Abu Taher, ,2019,Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection,2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), PP 643-646

[15] Kim Aechan, Mohyun Park, and Dong Hoon Lee. (2019) "AI-IDS: Application of Deep Learning to Real-Time Web Intrusion Detection." IEEE Access 8 (2019): 70245-70261.

[16] Kim, Kwangjo, Muhamad Erza Aminanto, and Harry Chandra (2018) “Network Intrusion Detection Using Deep Learning: A Feature Learning Approach.” Springer, 2018.

[17] Lee, Hoyeop, Youngju Kim, and Chang Ouk Kim. (2016) "A Deep Learning Model For Robust Wafer Fault Monitoring With Sensor Measurement Noise." IEEE Transactions on Semiconductor Manufacturing 30.1 (2016): 23-31.

[18] Manoj s. Koli, Manik K. Chavan, (2017) “An Advanced method for detection of botnet traffic using Internal Intrusion Detection”, 2017 International Conference on (ICICCT), March 10-11, 2017, Sangli, India.

[19] Ripon Patgiri, Udit Varshney, Tanya Akutota, and Rakesh Kunde,2018,An Investigation on Intrusion Detection System Using Machine Learning, IEEE Symposium Series on Computational Intelligence SSCI 2018, PP 1684-1691

[20] Shenfield, Alex, David Day, and Aladdin Ayesh. (2018) "Intelligent Intrusion Detection Systems Using Artificial Neural Networks." ICT Express 4.2 (2018): 95-99.

[21] Shengyi Pan, Thomas Morris, Uttam Adhikari, (2015) “Developing a Hybrid Intrusion Detection System using Data Mining for power system”, IEEE Transactions on, vol. 6, issues. 6, Nov. 2015.

[22] Skhumbuzo Zwane,Paul Tarwireyi, Matthew Adigun,2018, Performance Analysis of Machine Learning Classifiers for Intrusion Detection, IEEE

[23] Souparnika Jayaprakash Kamalanathan Kandasamy,2018,Database Intrusion Detection System Using Octraplet and Machine Learning,2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2, PP 1413-1416

[24] Vinayakumar, R.,(2018) "Deep learning approach for intelligent intrusion detection system." IEEE Access 7 (2018): 41525-41550.

[25] Yu-Lun Wan, Jen-Chun Chang,Rong-Jaye Chen,2018,Feature-Selection-Based Ransom-ware Detection with Machine Learning of Data Analysis,2018 3rd International Conference on Computer and Communication Systems, IEEE PP 85-88