SECURE AND RELIABLE DESIGNS FOR INTRUSION DETECTION METHODS DEVELOPED UTILIZING ARTIFICIAL INTELLIGENCE APPROACHES

Dr. Vinod Varma Vegesna
Sr. IT Security Risk Analyst, The Auto Club Group, United States of America.
Email: vinodvarmava@gmail.com

Abstract: Intrusion detection system plays an important role in network security. Intrusion detection model is a predictive model used to predict the network data traffic as normal or intrusion. Machine Learning algorithms are used to build accurate models for clustering, classification and prediction. In this paper classification and predictive models for intrusion detection are built by using machine learning classification algorithms namely Logistic Regression, Gaussian Naive Bayes, Support Vector Machine and Random Forest. These algorithms are tested with NSL-KDD data set. Experimental results show that Random Forest Classifier out performs the other methods in identifying whether the data traffic is normal or an attack.

Keywords: Classification Algorithms; Intrusion Detection; Machine Learning; Network Security; Supervised Learning, Data Analytics; Big Data; Anomaly based algorithm; Classification algorithms; Intrusion detection; Cyber Security; Cloud Security; Network; Cyber; Cyber Threats; Threat Analysis; Information Security; Data security.

1. Introduction

In the present era, it is difficult to envision world without internet. Every person has dependency on internet. It has become an important model in various applications such as education, business and others. So security of the data that is communicated through internet is necessary. Secure network is maintained by Intrusion Detection System (IDS). IDS observes the data traffic carefully and identifies it as normal or spam. Nowadays most of the applications depends on the advance network technologies namely wireless networks, wireless sensor networks and Bluetooth [1-9]. In case of wireless sensor networks security mechanisms such as key-management protocols, authentication techniques and security protocols cannot be used because of resource constraints. Intrusion Detection System is the ideal security mechanism for wireless sensor networks.

1.1.Intrusion Detection System

A security mechanism used to monitor the abnormal behavior of the network is an Intrusion Detection System (IDS) [10-17]. The IDS identifies and informs that whether the user activity is normal or not. The user's activities are compared by the IDS with the already stored intrusion records to identify the intrusion.

Accurate predictive models can be built for large data sets using supervised machine learning techniques that are not possible by traditional methods. As specified by Tom Mitchell, machine learning based intrusion detection falls under two categories Anomaly and Misuse. IDS learns the patterns by the training data, so the misuse based method is used. Misuse based detection can detect only the known attack, new attacks cannot be identified. Anomaly based IDS observes the normal behavior and if there is a change in the behavior then it considers that behavior as anomaly. So anomaly based IDS can detect new attacks that are not learned from the training model.

Till now different machine learning techniques such as artificial neural networks, Support Vector Machine and Naive Bayes based techniques are proposed for the intrusion detection. A new detection by combining different techniques, a hybrid detection technique is proposed. The literature on comparison of supervised machine learning techniques in intrusion detection is limited [18-27]. Hence this paper aims at understanding the

implications of using supervised machine learning techniques on intrusion detection.

The paper is organized as follows. Section 2 discusses some research work on Intrusion Detection System. In Section 3 various supervised machine learning methods used are discussed. Section 4 gives brief introduction about the data- set used. Section 5 discusses the methodology. Conclusion and future scope is given in section 7.

2. Related Work

Recently authors used algorithms namely Random Forest, Naive Bayes, K-means and Support Vector Machine to identify four types of attacks. They also proposed best feature selection method. Concluded that the Random Forest Classifier (RFC) outperforms the other methods. They have mentioned that hierarchical clustering method can be used to improve the performance. Authors proposed semi supervised machine learning based intrusion detection. Authors have not considered the resource consumption. Combination of different classifiers to identify the intrusion is proposed. They used supervised classification or unsupervised clustering for filtering of the data. They used NSL-KDD data-set and tested with decision tree classifier. But the proposed method works only for binary class classification.

Authors proposed intrusion detection system using supervised machine learning techniques to identify the on line network data as normal or not. The proposed method identifies probe and Denial of Service attacks only, but the other attacks are not considered.

A framework of machine learning approach is proposed. Intrusion is identified by analyzing the local features. Researchers proposed Naive Bayes based multiclass classifier to identify the intrusions. They suggested that intrusion detection is possible by Hidden Naive Bayes (HNB) model. Denial of Service attacks is identified with good accuracy compared to other attacks.

Studies suggested Intrusion detection technique using Support Vector Machine (SVM). They also used feature removal method to improve the efficiency. Using the proposed feature removal method they selected best nineteen features from the KDD-CUP99 data-set. In the proposed method the data set used is very small. A light weight IDS is proposed. The proposed method mainly focused on pre-processing of the data so that only important attributes can be used. The first step is to remove the redundant data so that the learning algorithms give the unbiased result.

A survey on intrusion detection systems was conducted. Information about IDSs such as classification, Intrusion type, computing location and infrastructure are discussed. They discussed about the Mobile Adhoc Networks (MANET) IDS. They compared MANETIDS and the Wireless Sensor Networks (WSN) IDS. Authors suggested that for mobile applications distributed and cooperative IDS schemes are suitable. For stationary applications centralized IDSs are suitable and for cluster based applications hierarchical IDSs are suitable. Authors proposed intrusion detection framework to detect routing attacks [28-36]. Specification based approach is used to detect routing attacks. Authors claim that the proposed method has low False Positive Rate (FPR) and good intrusion detection rate. The proposed method works only for static networks. Investigators developed IDS for Sink, Cluster Head (CH) and for a Sensor Node (SN) separately and combined altogether to identify the intrusion in heterogeneous Cluster Based Wireless Sensor Networks (CWSN) but the detection rate for U2R, R2L and Probe attacks is very low.

3. Supervised Machine Learning Techniques

Our work is to design a network intrusion detection system with the different supervised machine learning classifiers. This paper is to investigate the performance of the classifiers namely Logistic Regression ,Support Vector Machine(SVM), Gaussian Naive Bayes (GNB) and Random Forest in intrusion detection. These classifiers are discussed below.

3.1. Logistic regression

To solve the classification problems Logistic Regression(LR) is used. TLR works for both binary classification and multi class classification. Probability of occurrence of an event is predicted by fitting data to the Logistic function. The values selected by the logistic function are in the range 0 and 1. If the value is 0.5 and above then it is labeled asotherwise0.

3.2. Support vector machine

Mainly for classification problems Support Vector Machine algorithm is used, but it can be used in regression problems also. N-dimensional feature space is considered to plot each data item as a point with the value of each feature as a particular coordinate. Then classification is made by finding the hyper-plane that differentiates the two classes very well. Support

Vectors are the co-ordinates of specific observation that lies closest to the boarder line. In case of SVM training samples are divided into different subsets called as support vectors, the decision function is specified by these support vectors. This paper is based on the liner kernel method of SVM for intrusion detection.

3.3. Gaussian naive Bayes

The Gaussian Naive Bayes algorithm is the supervised learning method. Probabilities of each attribute which belongs to each class are considered for a prediction. This algorithm is assumes that the probability of each attribute belonging to a given class value is not depends on all other attributes. If the value of the attribute is known the probability of a class value is called as the conditional probabilities. Data instances provability can be found out by multiplying all attributes conditional probabilities together. Prediction can be made by calculating the each class instance probabilities and by selecting the highest probability class value.

3.4. Random forest classifier

In 2001 Breiman proposed the random forest machine learning classifier. It is a collaborative method which works based on the proximity search. It is decision tree based classifier. It makes use of standard divide and conquers approach to improve the performance. The main principle behind random forest is that strong learner group is created by a group of weak learners. It is applicable to disjunctive hypothesis. Comparison of Above mentioned classifiers based on the Precision, Recall,F1-Score and Accuracy is discussed insection 6.

4. Intrusion Detection Data-Set

The standard intrusion detection data set KDDCUP9923 has redundant records. This may lead to unfair result of the machine learning algorithms. So the supervised machine learning algorithms are tested NSL-KDD24 data-set which is the advanced version of the KDDCUP9 intrusion detection data-set. It has 42 features and the four simulated attacks [37-41].

Denial of Service (DoS) attack: Over usage of the bandwidth or non-availability of the system resources leads to the DoS attacks. Examples: Neptune, Teardrop and Smurf.
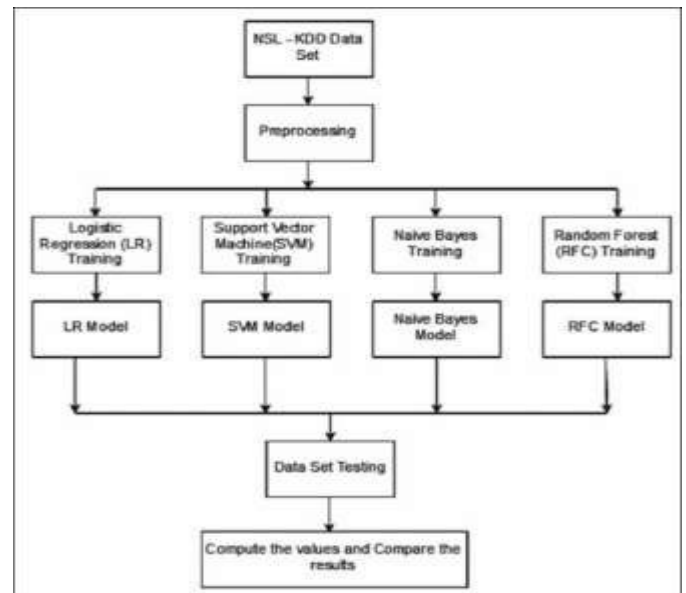
Fig.1. Methodology

User to Root (U2R) Attack: Initially attacker access normal user account, later gain access to the root by exploiting the vulnerabilities of the system. Examples: Perl, Load Module and Eject attacks.

Probe Attack: Have an access to entire network information before introducing an attack.
Examples: ipsweep, nmap attacks.

Root to Local (R2L) Attack: By exploiting some of the vulnerabilities of the network attacker gains local access by sending packets on a remote machine. Examples: imap, guess password and ftp- write attacks.

5. Methodology

The methodology used is shown in the Fig 1. In pre- processing step all the categorical data which are in textual form are converted to numerical form. Pre-processed data is divided as testing data and training data. The models are built using Logistic Regression, Gaussian Naïve Bayes, Support Vector Machine and Random Forest classifiers. These models are used for predicting the labels of the test data. Actual labels and predicted labels are compared. Accuracy, True Positive Rate (TPR) and False Positive Rate (FPR) are computed. Based on these parameters performance of the models are compared.
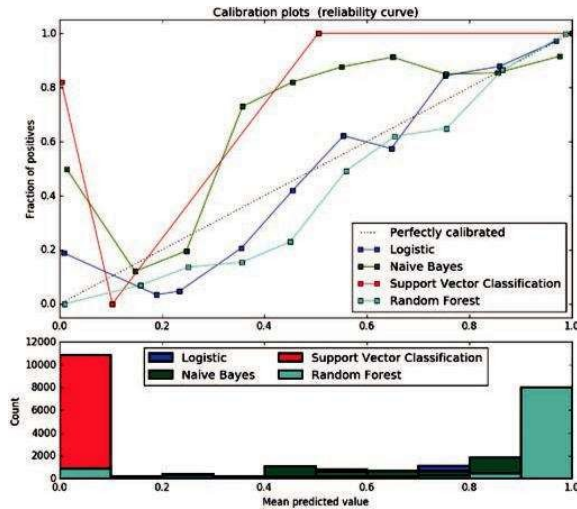
Fig. 2. Calibration Plots

Following steps are used to build the models.

- Pre-process the data set.
- The data set is divided as training data and testing data
- Build the classifier model on training data for
- Logistic Regression
- Support Vector Machine
- Gaussian Naive Bayes
- Random Forest
- Read the test data
- Test the classifier models on training data
- Compute and compare TPR, FPR, Precision, Recall, F1-Score and Accuracy for all the models.

6. Experimental Results and Discussions

Supervised machine learning algorithms namely Logistic Regression, Gaussian Naive Bayes, Support Vector Machine and Random Forest are tested on NSL-KDD dataset, the new standard intrusion detection data-set. These algorithms are tested on Intel Core (TM) i5- 3230M CPU @2.60 GHZ, 4 GB RAM and coding is done by Python. The result of the experiment is represented as a Reliability curve. In Reliability curve estimated probabilities are plotted against the true empirical probabilities. Figure 2 shows the Reliability Curve for the above mentioned supervised machine learning classifiers. Reliability curve for the ideal classifier falls near the diagonal because the estimated probabilities and empirical probabilities are nearly equal.

X-axis probability space is divided into ten bins as shown in Fig. 2. Estimated probabilities values ranging from 0 to 0.1, to 0.2 and so on. The values 0 to 0.1 belongs to I bin, 0.1 to 0.2 belongs to II bin and similarly the other ranges.

From the graph shown in Fig. 2, it can be concluded that the Random Forest classifier out performs the other methods in identifying the network traffic as normal or an attack. Whereas the SVM identifies the intrusion with the lowest probability estimate.

Quality of the classification models is identified by plotting the Receiver Operating Characteristics (ROC) curve. In ROC curve shows FPR verses TPR. ROC curve for the above mentioned classifiers is shown in the Fig. 3. Random Forest has highest TPR. Hence, the ROC curve for Random forest is plotted separately. By observing the graphs, it can be concluded that the Random forest classifier has lowest FPR and highest TPR in identifying attacks. It outperforms the other techniques. Whereas Support Vector Machine has highest FPR (39%) and minimal TPR (75%) for intrusion detection. This is due to the fact that too many features from the data set is considered15 and SVM's linear kernel function is used.
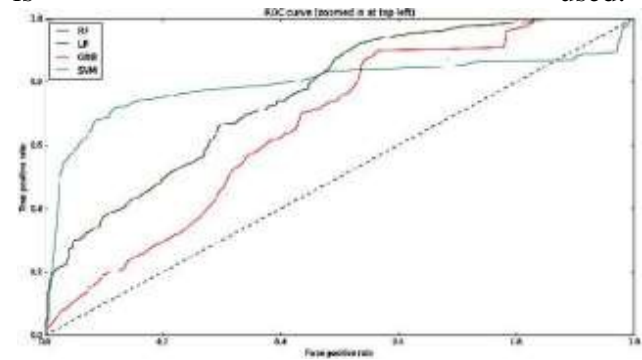
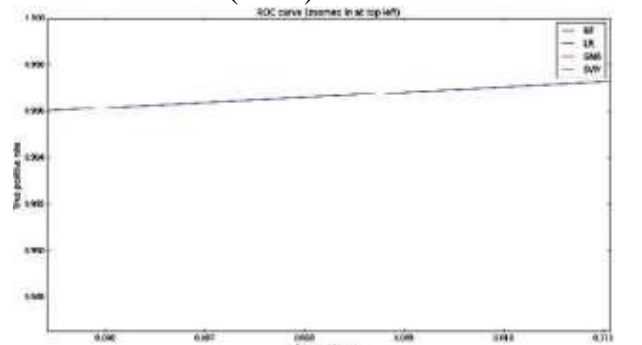

Fig.3(a). Receiver Operating Characteristics (ROC) curve

Fig.3(b).ROC for Random Forest Classifier

Table1. PerformanceMeasures

–	Precision	Recall	F1-Score	Accuracy
LR	0.83	0.85	0.82	0.84
GNB	0.79	0.81	0.78	0.79
SVM	0.76	0.79	0.77	0.75
RFC	0.99	0.99	0.99	0.99

Table 1 shows Precision, Recall, F1-Score and Accuracy of the supervised machine learning classifiers in identifying the intrusion. Based on the results shown in the Table 1 it can be identified that Random Forest classifier with the highest accuracy, outperforms the othermethods. Whereas SVM has the lowest accuracy, Logistic Regression algorithm has the good accuracy than Gaussian Naive Bayes and SVM.

7. Conclusions and Future Scope

An attempt has been made to check the performance of the supervised machine learning classifiers namely Support Vector Machine, Random Forest, Logistic Regression and Gaussian Naive Bayes are compared for intrusion detection. These algorithms are tested with the NSL-KDD data-set. Effective classifier is identified by comparing the performances basedon the precision, recall, F1-Score and accuracy. From the observed results it can be concluded that the Random forest classifier outperforms other classifiers for the considered data-set and parameters. It has the accuracy of 99%. The work can be extended by considering the classifiers for multiclass classification and considering only the important attributes for the intrusion detection..

References

[1] Zheng T, Ardolino M, Bacchetti A, Perona M (2021) The applications of Industry 4.0 technologies in manufacturing context: a systematic literature review. Int J Prod Res 59(6):1922–1954

[2] Ferrag MA, Shu L, Djallel H, Choo KKR (2021) Deep learning-based intrusion detection for distributed denial of service attack in agriculture 4.0. Electronics 10(11):1257

[3] Vinod Varma Vegesna (2022). "Investigations on Cybersecurity Challenges and Mitigation Strategies in Intelligent transport systems," Irish Interdisciplinary Journal of Science and Research, Vol. 6, Iss. 4, Pages 70-86, October-December 2022, doi: 10.46759/iijsr.2022.6409.

[4] Kiran MB (2021) Significance of intruder detection techniques in the context of industry 4.0. In: Proceedings of the international conference on industrial engineering and operations management. pp 2977–2985

[5] Gunduz MZ, Das R (2020) Cyber-security on smart grid: Threats and potential solutions. Comput Netw 169:107094

[6] Hamid Ali Abed Al-Asadi and Majda Ali Abed, "Object Recognition Using Artificial Fish Swarm Algorithm on Fourier Descriptors," American Journal of Engineering, Technology and Society; Volume 2, Issue 5: pp. 105-110, 2015.

[7] Ahmad I, Shah SAA, Al-Khasawneh MA (2021) Performance Analysis of Intrusion Detection systems for smartphone security enhancements. In: 2021 2nd international conference on smart computing and electronic enterprise (ICSCEE), pp 19–25, IEEE

[8] Sun M, Li X, Yang R, Zhang Y, Zhang L, Song Z, Liu Q, Zhao D (2020) Comprehensive partitions and different strategies based on ecological security and economic development in Guizhou Province. China J Clean Prod 274:122794

[9] Vinod Varma Vegesna (2022). "Methodologies for Enhancing Data Integrity and Security in Distributed Cloud Computing with Techniques to Implement Security Solutions," Asian Journal of Applied Science and Technology, Volume 6, Issue 2, Pages 167-180, April-June 2022, doi: 10.38177/ajast.2022.6217.

[10] Saghezchi FB, Mantas G, Violas MA, de Oliveira Duarte AM, Rodriguez J (2022) Machine learning for DDoS attack detection in industry 4.0 CPPSs. Electronics 11(4):602

[11] Saxena N, Hayes E, Bertino E, Ojo PP, Choo KKR, Burnap PP (2020) Impact and key challenges of insider threats on organizations and critical businesses. Electronics 9(9):1460

[12] Alohali MA, Al-Wesabi FN, Hilal AM, Goel S, Gupta D, Khanna A (2022) Artificial intelligence enabled intrusion detection systems for cognitive cyber-physical systems in industry 4.0 environment. Cognit Neurodyn 16:1–13

[13] Tahir B, Jolfaei A, Tariq M (2021) Experience driven attack design and federated learning based intrusion detection in industry 4.0. IEEE Trans Ind Inf 18:6398–6405

[14] Vinod Varma Vegesna (2021). "Analysis of Data Confidentiality Methods in Cloud Computing for Attaining Enhanced Security in Cloud Storage," Middle East Journal of Applied Science & Technology, Vol. 4, Iss. 2, Pages 163-178, April-June 2021, Available at SSRN: https://ssrn.com/abstract=4418127

[15] Qi L, Yang Y, Zhou X, Rafique W, Ma J (2021) Fast anomaly identification based on multi-aspect data streams for intelligent

intrusion detection toward secure industry 4.0. IEEE Trans Ind Inf 18:6503–6511

[16] Vinod Varma Vegesna (2020). "Secure and Privacy-Based Data Sharing Approaches in Cloud Computing for Healthcare Applications," Mediterranean Journal of Basic and Applied Sciences, Volume 4, Issue 4, Pages 194-209, October-December 2020, doi: 10.46382/mjbas.2020.4409.

[17] Hamzah F. Zmezm, HarethZmezm, Mustafa S.Khalefa, Hamid Ali Abed Al-Asadi, "A Novel Scan2Pass Architecture for Enhancing Security towards E-Commerce," Future Technologies Conference 2017, 29-30 November 2017 | Vancouver, BC, Canada, 2017.

[18] Yang K, Shi Y, Yu Z, Yang Q, Sangaiah AK, Zeng H (2022) Stacked one-class broad learning system for intrusion detection in industry 4.0. IEEE Trans Ind Inf 19:251–260

[19] Ibitoye O, Shafiq O, Matrawy A (2019) Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks. In: 2019 IEEE global communications conference (GLOBECOM), pp. 1–6. IEEE

[20] Vinod Varma Vegesna (2019). "Investigations on Different Security Techniques for Data Protection in Cloud Computing using Cryptography Schemes", Indo-Iranian Journal of Scientific Research, Volume 3, Issue 1, Pages 69-84, January-March 2019, Available at SSRN: https://ssrn.com/abstract=4418119

[21] Saghezchi FB, Mantas G, Violas MA, de Oliveira Duarte AM, Rodriguez J (2022) Machine learning for DDoS attack detection in industry 4.0 CPPSs. Electronics 11(4):602

[22] HarethZmezm, Mustafa S.Khalefa, Hamid Ali Abed Al-Asadi, Hamzah F. Zmezm, Dr. Hussain Falih Mahdi, Hassan MuhsenAbdulkareem Al-Haidari. "Suggested Mechanisms for Understanding the Ideas in Authentication System," International Journal of Advancements in Computing Technology9(3):10-24, 2018.

[23] Khoa TV, Saputra YM, Hoang DT, Trung NL, Nguyen D, Ha NV, Dutkiewicz E (2020) Collaborative learning model for cyberattack detection systems in iot industry 4.0. In: 2020 IEEE wireless communications and networking conference WCNC, pp. 1–6. IEEE.

[24]Hamid Ali Abed Al-Asadi and et., "Nature Inspired Algorithms multi-objective histogram equalization for Grey image enhancement", Advances in Computer, Signals and Systems (2020) 4: 36-46 Clausius Scientific Press, Canada DOI: 10.23977/acss.2020.040106.

[25] Li B, Wu Y, Song J, Lu R, Li T, Zhao L (2020) DeepFed: Federated deep learning for intrusion detection in industrial cyber–physical systems. IEEE Trans Industr Inf 17(8):5615–5624

[26] Chowdhury D, Poddar S, Banarjee S, Pal R, Gani A, Ellis C, Arya RC, Gill SS, Uhlig S (2022) CovidXAI: explainable ai-assisted web application for COVID-19 vaccine prioritisation. Int Technol Lett. https://doi.org/10.1002/itl2.381pp.e381

[27] Vinod Varma Vegesna (2018). "Analysis of Artificial Intelligence Techniques for Network Intrusion Detection and Intrusion Prevention for Enhanced User Privacy", Asian Journal of Applied Science and Technology, Volume 2, Issue 4, Pages 315-330, Oct-Dec 2018, Available at SSRN: https://ssrn.com/abstract=4418114

[28] Krishnaveni S, Sivamohan S, Sridhar S, Prabhakaran S (2022) Network intrusion detection based on ensemble classification and feature selection method for cloud computing. Concurr Comput Pract Exp 34(11):e6838

[29] Krishnaveni S, Sivamohan S, Sridhar SS, Prabakaran S (2021) Efficient feature selection and classification through ensemble method for network intrusion detection on cloud computing. Clust Comput 24(3):1761–1779

[30] Vinod Varma Vegesna (2017). "Incorporating Wireless Sensor Networks and the Internet of Things: A Hierarchical and Security-Based Analysis," International Journal of Current Engineering and Scientific Research, Volume-4, Issue-5, Pages 94-106, Available at SSRN: https://ssrn.com/abstract=4418110

[31] Barnard PP, Marchetti N, DaSilva LA (2022) Robust network intrusion detection through explainable artificial intelligence (XAI). IEEE Netw Lett 4(3):167–171

[32] Hamid Ali Abed Al-Asadi and et al., "Critical Comparative Review of Nature-Inspired Optimization Algorithms (NIOAs), International Journal of Simulation: Systems, Science and Technology (IJSSST), 2020, 21(3), PP1-15

[33] Liu H, Zhong C, Alnusair A, Islam SR (2021) FAIXID: a framework for enhancing aiexplainability of intrusion detection results using data cleaning techniques. J Netw Syst Manage 29(4):1–30

[34] Vinod Varma Vegesna (2016). "Threat and Risk Assessment Techniques and Mitigation Approaches for Enhancing Security in Automotive Domain," International Journal of Management, Technology And Engineering, Volume VI, Issue II, July-Dec 2016, Pages 314-331, Available at SSRN: https://ssrn.com/abstract=4418100

[35] Larriva-Novo X, Villagrá VA, Vega-Barbas M, Rivera D, Sanz Rodrigo M (2021) An IoT-focused intrusion detection system approach based on preprocessing characterization for cybersecurity datasets. Sensors 21(2):656

[36] Li X, Yi PP, Wei W, Jiang Y, Tian L (2021) LNNLS-KH: a feature selection method for network intrusion detection. Sec Commun Netw. https://doi.org/10.1155/2021/8830431

[37] Vinod Varma Vegesna (2015). "Incorporating Data Mining Approaches and Knowledge Discovery Process to Cloud Computing for Maximizing Security," International Journal of Current Engineering and Scientific Research, Volume-2, Issue-6, Pages 118-133, Available at SSRN: https://ssrn.com/abstract=4418107

[38] Resma KB, Nair MS (2021) Multilevel thresholding for image segmentation using Krill Herd optimization algorithm. J King Saud Univ-Comput Inf Sci 33(5):528–541

[39] Abdul W, Alsulaiman M, Amin SU, Faisal M, Muhammad G, Albogamy FR, Bencherif MA, Ghaleb H (2021) Intelligent real-time Arabic sign language classification using attention-based inception and BiLSTM. Comput Electr Eng 95:107395

[40] Knapič S, Malhi A, Saluja R, Främling K (2021) Explainable artificial intelligence for human decision support system in the medical domain. Mach Learn Knowl Extractio 3(3):740–770

[41] Kwon D, Natarajan K, Suh SC, Kim H, Kim J (2018) An empirical study on network anomaly detection using convolutional neural networks. In: ICDCS pp 1595–1598