Machine Learning for Classification analysis of Intrusion Detection on NSL-KDD Dataset

Faheem Masoodi[1], Alwi M Bamhdi[2], Tawseef A Teli[3]

[1]University of Kashmir, J&K India
[2]Umm Ul Qura University, KSA
[3]Amarsingh College, J&K India

Article History Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract. In the existing digital era, security concerns turn out to be a prime obstacle as it hampers the user's privacy. Moreover, with the emergence of new technologies, enormous amount of data is present on the network which is subjected to innumerable malicious attacks and security vulnerabilities. It is therefore, essential to detect these vulnerabilities on time so that the privacy of the user's data is not hampered and for the same intrusion detection system (IDS) is used which is deemed as the cornerstone of security. An IDS is indispensable for timely detection of cyber-attacks as it is capable of detecting intrusive activities adequately so that any potential harm to the system resources and the user base can be avoided in time. Owing to their understanding of IDS for reducing security threats, in the current work, machine learning classifiers (MLC) were used for classifying the data. The system performance was evaluated using four diverse attribute subsets obtained from NSL-KDD dataset. For optimization, prior to the training/testing phase, the dataset was pre-processed so that irrelevant features could be removed as their contribution is inconsequential in detecting attack classes. Finally, the overall model accuracy for different attack classes namely DoS, Probe, U2L, and R2L was compared to detect the most suitable algorithm for a particular attack class

Keywords: Intrusion detection, classification, Machine learning

1. Introduction

Although numerous security mechanisms like encryption techniques and firewalls have been deployed over the time, yet it is impossible to avoid attacks, as security breaches are witnessing a spike (Siddiqui et al., 2019). Owing to these facts, security and privacy has become a prime area of research, as there is a need to develop a mechanism that can proactively detect attacks (Masoodi et al., 2019). To address these threats, it is essential to have an IDS that can adapt to the security policies, allowing real-time corrective mechanism without causing any severe damage to the user or system (Ahmed & Masoodi, 2020). IDS is a reliable and effective means of detecting unwanted attempts of accessing, manipulating, or disabling a computer system, primarily via internet (Alam et al., 2019). It scrutinizes network traffic for detecting malicious activities that can pose a threat to the network or system and thus, is a critical means of defence for network controllers. IDS are broadly categorized as misuse (MIDS) and anomaly detection (AIDS) (Lv et al., 2020). In MIDS, data gathered from the network is compared with a predefined attack signatures present in the database to determine an attack class (Alazzam et al., 2020). Although MIDS has a significant detection rate, they are not able to detect new attack classes. On the contrary, in AIDS, anything that deviates from the normal activities is marked as an attack.

With the advent of data analysis technology, IDS has incessantly evolved. They have become more reliable and powerful as the initial statistical theory has been replaced by artificial intelligence algorithms for the detection of attack classes. However, owing to the complex and diverse behavior of intrusive activities, the existing methodologies have certain loopholes that must be addressed on time to prevent any severe damage. Facing the rise in security issues, researchers have developed new intrusion detection techniques based on the merits of existing ML techniques like decision tree (Ferrag et al., 2020), (Rajagopal et al., 2020), K-Nearest neighbor (KNN)(Prasad et al., 2020) , random forest (RF) (W. Wang et al., 2018),(Ambikavathi & Srivatsa, 2020) and the same have been investigated in the next section to present a brief overview. However, in most of the research works, only the final results have been highlighted whereas in the current work an inclusive and comprehensible insight of the work has been presented systematically. The rationale of the current research was to present a comparative analysis of various MLC for intrusion detection, which is competent, resourceful, and robust.

2. Overview

An intrusion detection system works as a mechanism to identify the several attacks in a network and a comprehensive review of the work performed on IDS has been described in this section. Machine learning procedures are widely applied to create robust IDS as these approaches have the capacity to handle huge volume of data and consequently network authorities can apply more powerful mechanism to prevent the information system (Shuaib et al., 2019).

Feature selection is key component of intrusion detection system as it selects important features from the dataset so that the accuracy rate of the model can be increased. In (Alazzam et al., 2020), feature selection method based on wrapper approach was developed with the use pigeon inspired optimizer, which selects informative attributes from the set of features. The described approach was examined on UNSW-NB15, KDDCUP 99 and NLS-KDD. Similarly, in (Kunhare et al., 2020), RF classification algorithm was used to remove less important features from the dataset to reduce the computational cost and then, various ML classifiers were used, namely SVM, KNN, logistic regression (LR), DT and Naïve Bayes (NB) to train and test the model. Finally PSO technique was used for enhancing the performance of the model, the model achieves an accuracy of 99% on NSL-KDD dataset, when 10 important attributes were selected for the classification purpose.

SVM has been widely used to detect intrusion in the network, (Bachar et al., 2020) employed SVM for the detection of intrusions and the model was examined on UNSW dataset. Accordingly, a comparable review of some machine learning classifiers e.g. MLP, RepTree and RF was directed, the model provides an classification rate of 94%. In (D. Wang & Xu, 2020) 3 updated types of SVM were employed in which 99.86% accuracy was achieved in terms of enhanced whale optimized SVM. In (Sumaiya Thaseen et al., 2020) , several machine learning classification algorithms were employed to for training as well as testing the model, in which a tool namely wireshark were utilized to collect network traffic i.e. packet data. Initially the data was pre-processed and classification rate of 83.6%, 95.1%, 98.2%, 99.8% was achieved using ML methods like NB, KNN SVM, and RF. Machine Learning and knowledge-based hybrid approach was employed to find out several categories of intrusions using KDD-99 (Peddabachigari et al., 2007). In the presented method, knowledge-based technique was utilized to traverse all the output classes to select the satisfactory model, from which the predictions were performed, and better results were obtained.

In (Sun et al., 2019), three machine learning algorithms, namely MLP, NB and SVM were examined employing several feature sets selected from original dataset. The outcome obtained from the experiment showed that the, MLP performs better than the NB and SVM classifier in detecting mixed, normal traffic and malicious windows. The proposed model could be used effectively as it fast and could announce alert to sophisticated intrusion detection system, the accuracy and FAR of the model is 92.09% and 0.27% respectively. Likewise, (Pham et al., 2018) employed neural network and MLP for inspecting intrusion in the network, in which the model was trained on the data that was classified as normal and attack i.e. two classes. In case of, MLP back propagation method was utilized in testing phase, KDD-99 dataset was employed to verify model performance. With all features the proposed model obtained an accuracy of 94% and 91% with 2 hidden layers.

In (Ghazy et al., 2020) , various variable selection approaches and ranking approaches were used for detecting intrusion. Appropriately, the RF classifier was trained, tested by using NSL-KDD. Better results were achieved using feature selection technique based on wrapper method for probe and DoS attacks on 41 attributes and for U2R attack on 11 attributes were selected. Identically, in (Waskle et al., 2020), RF and PCA (principal component analysis) were used to detect the intrusions where the PCA was employed to reduce the of the data, and RF was employed to categorize the data as malicious or normal. A relative study with several other machine learning algorithm like SVM, DT and NB was directed and classification rate of 96.78% was achieved. In (Nancy et al., 2020) , an IDS was developed by using DT-based approach and the model performance were examined on two datasets namely on BOT-IoT and CICIDS, in which an classification rate of 96.995% and 96.665% was obtained, accordingly. Similarly, in (Ahmad et al., 2018), decision tree technique with fuzzy rules and temporal were employed for training/testing given model using dataset KDD-99, in which dynamic recursive was used for variable selection was applied to select the important features for feature selection. The model obtained an accuracy of 99.99%, 95.23% , 92.67%, 57.39% for DoS, R2L, Probe, U2R respectively.

3. **Methodology**

The results produced by the presented model have been explained in current segment here KNN SVM, LR, MLP, NB, ETC, DT and RF were employed to classify the data instances as intrusive or normal. The performance of the model was evaluated on four randomly selected attribute subsets that where obtained from the original dataset namely NSL-KDD. The following Succeeding steps will outline and discuss exactly the whole process of the current work.

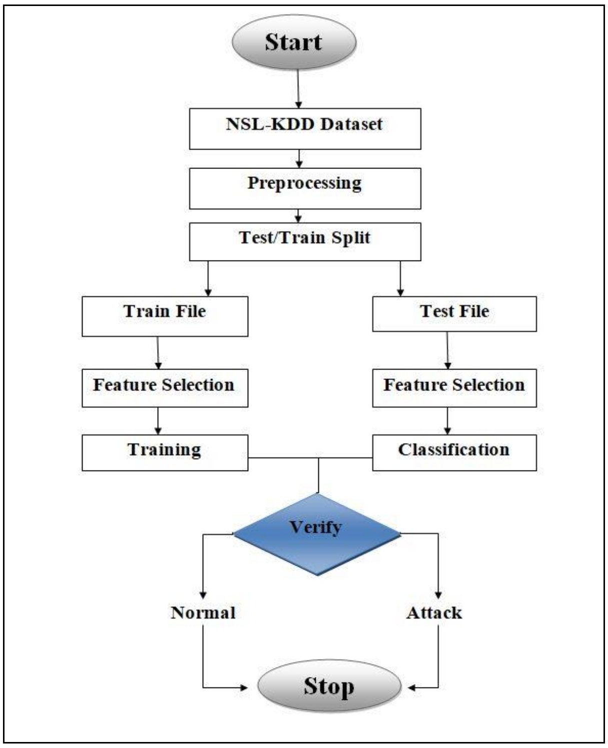


Figure I. Framework of the proposed model

NSL-KDD Dataset

In this approach, NSL-KDD dataset was used for performance analysis of the model. The results produced by the model were analyzed accordingly. During the training/testing, the instances selected randomly for the training phase consists of 189, 2836, 10431, 40047 and 57883 and the test dataset consists of 63, 913, 3646, 13337 and 19170 records of U2R, R2L, Probe, DoS and Normal. The distribution of NSL-KDD instances in various attack classes is shown in Figure II below.

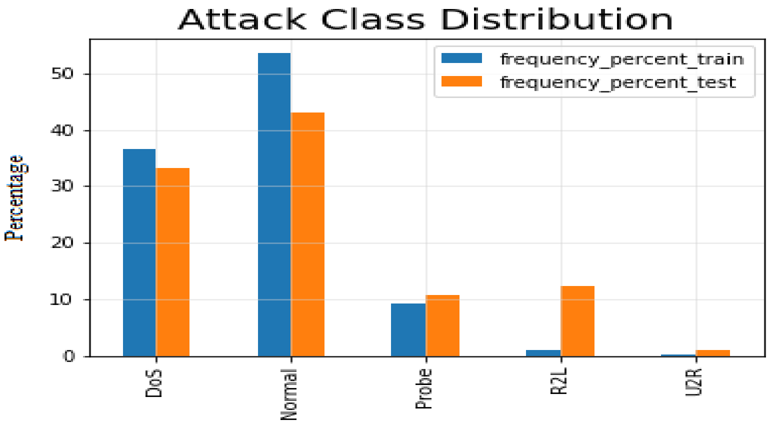
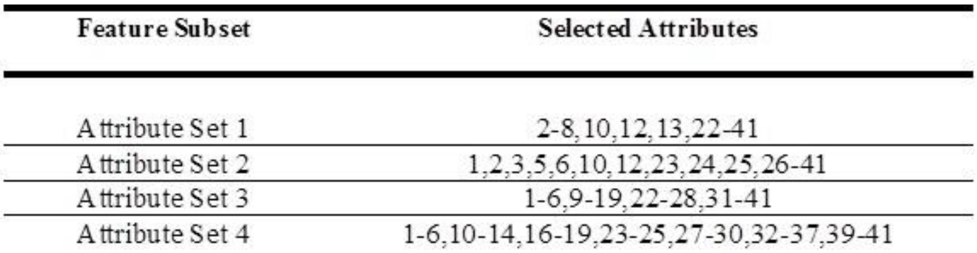


Figure II. Distribution of various attcks in the NSL-KDD Dataset

Preprocessing plays an important role through which non-numeric/categorical features are taken out or replaced by using certain feature extraction/selection techniques. Preprocessing decreases the computational intricacy and increases the classification rate of the model, in present work, four attribute subsets selected from the original feature set are presented in table 1 were selected for training and testing the proposed model. Since the appropriate features were selected to train, test the model, which reduces the training time notably. Therefore preprocessing is important task to minimize the dimensions of the original feature subset and to reduce the computational overheads during the processing of data.

Table 1. FOUR FEATURE SUBSETS USED FOR CLASSIFICATION



Feature Subset	Selected Attributes
Attribute Set 1	2-8,10,12,13,22-41
Attribute Set 2	1,2,3,5,6,10,12,23,24,25,26-41
Attribute Set 3	1-6,9-19,22-28,31-41
Attribute Set 4	1-6,10-14,16-19,23-25,27-30,32-37,39-41

Classification

A number of classification algorithms like Support vector machine, K- nearest neighbor, Decision trees, Logistic regression and Random Forest were used as intrusion inspection engine to categorize the data as intrusive or normal. Support vector machines can be used for both linear as well as non-linear data as it uses binary-class technique (Abrar et al., 2020). The goal of SVM is to map the input vector or extreme points, and then best hyper-plane is generated which is the optimal decision line that separates n-dimensional space into sub-spaces depends upon the highest segregation between the extreme points/support vectors. These support vectors are used to obtain the optimal hyper-plane by the SVM without using the entire training instances. Thus SVM powerful technique in detecting outliers (Ahmad et al., 2018). In the current work, a function known as radial basis kernel was employed to train/test the model; using kernel function for non-linear data provides better results. Like SVM, KNN belongs to the class of supervised learning algorithm in which the classifier uses Euclidean distance and a parameter K (define the number of nearest neighbors) to categorize the data according to the classes defined in the training data. In the current work, the value of K was assigned a value of five, so that the performance of model can be increased by fixing the actual location of the data point. Same process was repeated till the convergence and as a result the data is categorized into different classes (Li et al., 2014). LR is a supervised learning classification approach, which can be applied to determine the possible class of the output variable. The proposed model was trained and evaluated using a single hidden layer of 100 neurons and a rectified linear unit function. The classification problems are solved using a decision tree, which has a tree-like structure. The decision tree's nodes represent the features/attributes, while the paths represent the features/attributes' potential values (Peddabachigari et al., 2007). Gini index or information gain criteria can be used to split nodes, and in this study, Gini index was used to pick the best feature that effectively splits the data. DT uses recursive process to classify each data instance according to the values they contain and stops when the depth of tree reaches to none (Yao et al., 2006) . Random forest uses ensemble technique, it takes the outcome of multiple DTs in order to make the better prediction or classify the data according to their respective classes (Bamhdi et al., 2021). In the case of RF, important features are chosen at random from the original collection of attributes, providing clearer knowledge about data classification and allowing for the creation of an accurate classification model (Zhou et al., 2016).

Results

In the current approach, four feature subsets were selected randomly from the original dataset in an attempt to minimize the dimensions of the data to train and test the proposed model. Feature selection using random selection of features from original dataset is a successful method to decrease the model intricacy and training time. By analyzing the results produced by the model, this approach operates efficiently in the proposed model however; it may not be favorable in certain scenarios. A number of classification algorithms namely: KNN, SVM, NB, LR, MLP, ETC, DT, and RF were trained by using 1,11,386 instances and were tested on 37,129 samples of NSL-KDD dataset and the behaviour of these models were evaluated using different parameters. The results produced by these classifiers on different several feature subsets are accuracy, *F1-score, precision, and recall* which are graphically depicted in figures below.

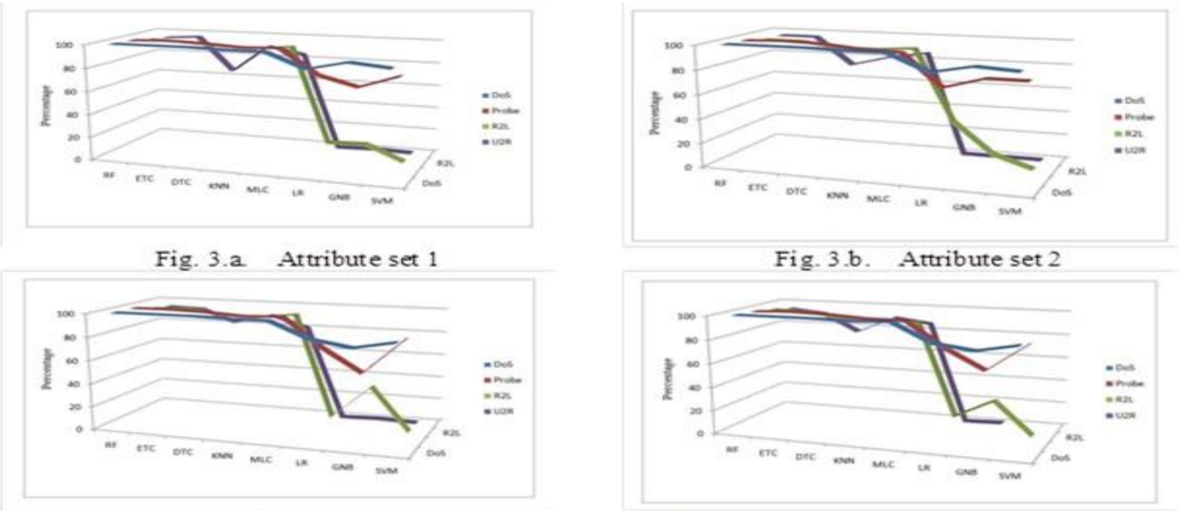


Fig. 3.a Attribute set 1

Fig. 3.b. Attribute set 2

Fig. 3.c. Attribute set 3

Fig. 3.d. Attribute set 4

Figure III. Graphical representation of **Precision** using various feature subsets

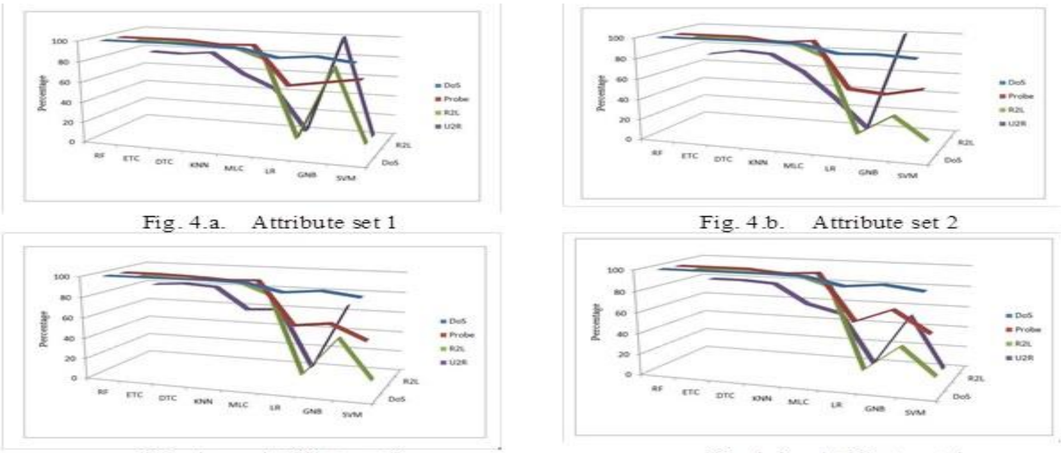


Fig. 4.a. Attribute set 1

Fig. 4.b. Attribute set 2

Fig. 4.c. Attribute set 3

Fig. 4.d. Attribute set 4

Figure IV. Graphical representation of **Recall** using various feature subsets

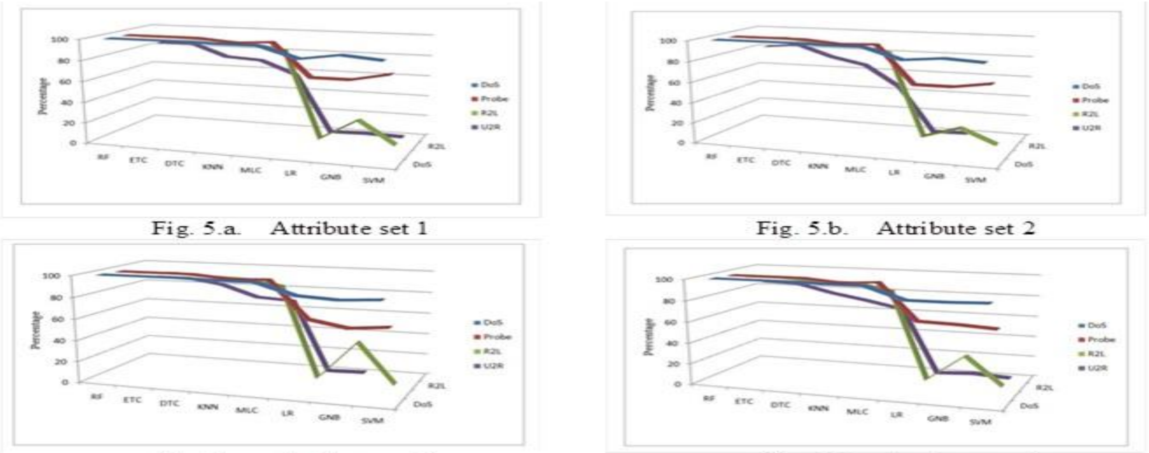


Fig. 5.a. Attribute set 1

Fig. 5.b. Attribute set 2

Fig. 5.c. Attribute set 3

Fig. 5.d Attribute set 4

Figure V. Graphical representation of **F1- score** using various feature subsets

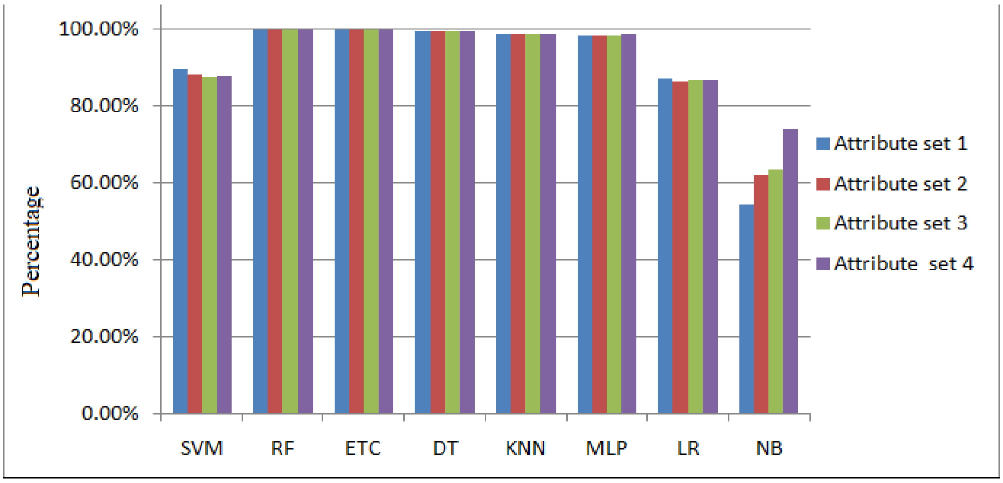


Figure VI. Model accuracy of various classifiers

Among the four types of attacks, promising results were obtained for DoS attacks because the training dataset contained fair examples of DoS attacks, while U2L attacks yielded insufficient results. The overall classification rate achieved by the model using RF, decision tree, and extra-tree classifiers on all four randomly selected feature subsets is greater than 99 percent, indicating that these classifiers are capable of handling inconsistent disseminate data. Random forests can detect data imbalances and, since it employs bootstrap techniques to improve minority category life, it reduces data classification errors, resulting in an increase in classification rate. The decision tree has a higher generalization capability and conducts a thorough examination of all of its potential branches, resulting in excellent results. Furthermore, these ML classifiers concentrate on cost bounding, allowing them to be excellent intrusion detectors. As a result, these machine learning classifiers can be used to detect intrusions quickly and effectively.

4. Conclusion and Future Scope

In the current work, a comparative study of performance analysis of various ML algorithms for intrusion detection on NSL-KDD dataset was conducted. For enhancing the training speed and reducing the computational intricacies of the model, only the relevant features were selected. Promising results were attained for various attack classes and the same has been presented systematically for a comprehensible insight of the current work. It can be concluded that a reliable IDS capable of detecting intrusion in real-time can be formed using various classifiers and accordingly unauthorized access to the network resources can be avoided.

Furthermore, as attackers use various vulnerabilities present in the machine learning algorithms to carry out attacks, this problem can be tackled in future by using an ensemble approach where a variety of classifiers are used in combination. Besides this, innovative feature selection methods can also be used to choose essential features from dataset, and accordingly, the performance of the model can be enhanced.

References

A. Abrar, I., Ayub, Z., Masoodi, F., & Bamhdi, A. M. (2020). A Machine Learning Approach for Intrusion Detection System on NSL-KDD Dataset. *Proceedings - International Conference on Smart Electronics and Communication, ICOSEC 2020*, 919–924. https://doi.org/10.1109/ICOSEC49089.2020.9215232

B. Ahmad, I., Basheri, M., Iqbal, M. J., & Rahim, A. (2018). Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection. *IEEE Access*, *6*, 33789–33795. https://doi.org/10.1109/ACCESS.2018.2841987

C. Ahmed, T., & Masoodi, F. (2020). *Security Concerns and Privacy Preservation in Blockchain based IoT Systems : Opportunities and Challenges*. *Icicnis*, 29–36.

D. Alam, S., Shuaib, M., & Samad, A. (2019). A Collaborative Study of Intrusion Detection and Prevention Techniques in Cloud Computing. In *Lecture Notes in Networks and Systems* (Vol. 55, pp. 231–240). https://doi.org/10.1007/978-981-13-2324-9_23

E. Alazzam, H., Sharieh, A., & Sabri, K. E. (2020). A feature selection algorithm for intrusion detection system based on Pigeon Inspired Optimizer. *Expert Systems with Applications*, *148*, 113249.

https://doi.org/10.1016/j.eswa.2020.113249

F. Ambikavathi, C., & Srivatsa, S. K. (2020). Predictor Selection and Attack Classification using Random Forest for Intrusion Detection. *Journal of Scientific and Industrial Research (JSIR)*, *79*(05), 365–368.

G. Bachar, A., Makhfi, N. El, & Bannay, O. El. (2020). Towards a behavioral network intrusion detection system based on the SVM model. *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology, IRASET 2020*. https://doi.org/10.1109/IRASET48871.2020.9092094

H. Bamhdi, A. M., Abrar, I., & Masoodi, F. (2021). An ensemble based approach for effective intrusion detection using majority voting. *Telkomnika (Telecommunication Computing Electronics and Control)*, *19*(2), 664–671. https://doi.org/10.12928/TELKOMNIKA.v19i2.18325

I. Ferrag, M. A., Maglaras, L., Ahmim, A., Derdour, M., & Janicke, H. (2020). RDTIDS: Rules and decision tree-based intrusion detection system for internet-of-things networks. *Future Internet*, *12*(3), 1–14. https://doi.org/10.3390/fi12030044

J. Ghazy, R. A., El-Rabaie, E. S. M., Dessouky, M. I., El-Fishawy, N. A., & El-Samie, F. E. A. (2020). Feature Selection Ranking and Subset-Based Techniques with Different Classifiers for Intrusion Detection. *Wireless Personal Communications*, *111*(1), 375–393. https://doi.org/10.1007/s11277-019-06864-3

K. Kunhare, N., Tiwari, R., & Dhar, J. (2020). Particle swarm optimization and feature selection for intrusion detection system. *Sādhanā*, *0123456789*. https://doi.org/10.1007/s12046-020-1308-5

L. Li, W., Yi, P., Wu, Y., Pan, L., & Li, J. (2014). A new intrusion detection system based on KNN classification algorithm in wireless sensor network. *Journal of Electrical and Computer Engineering*, *2014*(1). https://doi.org/10.1155/2014/240217

M. Lv, L., Wang, W., Zhang, Z., & Liu, X. (2020). A novel intrusion detection system based on an optimal hybrid kernel extreme learning machine. *Knowledge-Based Systems*, *195*, 105648. https://doi.org/10.1016/j.knosys.2020.105648

N. Masoodi, F., Alam, S., & Siddiqui, S. T. (2019). Security and privacy threats, attacks and countermeasures in Internet of Things. *Int. J. Netw. Secur. Appl*, 67–77.

O. Nancy, P., Muthurajkumar, S., Ganapathy, S., Santhosh Kumar, S. V. N., Selvi, M., & Arputharaj, K. (2020). Intrusion detection using dynamic feature selection and fuzzy temporal decision tree classification for wireless sensor networks. *IET Communications*, *14*(5), 888–895. https://doi.org/10.1049/iet-com.2019.0172

P. Peddabachigari, S., Abraham, A., Grosan, C., & Thomas, J. (2007). Modeling intrusion detection system using hybrid intelligent systems. *Journal of Network and Computer Applications*, *30*(1), 114–132. https://doi.org/10.1016/j.jnca.2005.06.003

Q. Pham, N. T., Foo, E., Suriadi, S., Jeffrey, H., & Lahza, H. F. M. (2018). Improving performance of intrusion detection system using ensemble methods and feature selection. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3167918.3167951

R. Prasad, T. B., Prasad, P. S., & Kumar, K. P. (2020). *An Intrusion Detection System Software Program Using K-NN Nearest Neighbours Approach*. *1*(1), 1–6.

S. Rajagopal, S., Kundapur, P. P., & Hareesha, K. S. (2020). A Stacking Ensemble for Network Intrusion Detection Using Heterogeneous Datasets. *Security and Communication Networks*, *2020*. https://doi.org/10.1155/2020/4586875

T. Shuaib, M., Samad, A., Alam, S., & Siddiqui, S. T. (2019). Why Adopting Cloud Is Still a Challenge?—A Review on Issues and Challenges for Cloud Migration in Organizations. In *Advances in Intelligent Systems and Computing* (Vol. 904, pp. 387–399). https://doi.org/10.1007/978-981-13-5934-7_35

U. Siddiqui, S. T., Alam, S., Shuaib, M., & Gupta, A. (2019). Cloud Computing Security using Blockchain. *Journal of Emerging Technologies and Innovative Research*, *6*(6), 791–794.

V. Sumaiya Thaseen, I., Poorva, B., & Ushasree, P. S. (2020). Network Intrusion Detection using Machine Learning Techniques. *International Conference on Emerging Trends in Information Technology and Engineering, Ic-ETITE 2020*. https://doi.org/10.1109/ic-ETITE47903.2020.148

W. Sun, L., Ho, A., Xia, Z., Chen, J., & Zhang, M. (2019). Development of an early warning system for network intrusion detection using benford’s law features. In *Communications in Computer and Information Science: Vol. 1095 CCIS*. Springer Singapore. https://doi.org/10.1007/978-981-15-0758-8_5

X. Wang, D., & Xu, G. (2020). Research on the detection of network intrusion prevention with SVM based optimization algorithm. *Informatica (Slovenia)*, *44*(2), 269–273. https://doi.org/10.31449/inf.v44i2.3195

Y. Wang, W., Li, Y., Wang, X., Liu, J., & Zhang, X. (2018). Detecting Android malicious apps and categorizing benign apps with ensemble of classifiers. *Future Generation Computer Systems*, *78*, 987–994. https://doi.org/10.1016/j.future.2017.01.019

Z. Waskle, S., Parashar, L., & Singh, U. (2020). Intrusion Detection System Using PCA with Random Forest Approach. *Proceedings of the International Conference on Electronics and Sustainable*

Communication Systems, *ICESC 2020*, 803–808. https://doi.org/10.1109/ICESC48915.2020.9155656

AA. Yao, Y., Yang, W., Gao, F. X., & Yu, G. (2006). Anomaly intrusion detection approach using hybrid MLP/CNN neural network. *Proceedings - ISDA 2006: Sixth International Conference on Intelligent Systems Design and Applications*, *2*, 1095–1102. https://doi.org/10.1109/ISDA.2006.253765

BB. Zhou, Q., Zhou, H., & Li, T. (2016). Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features. *Knowledge-Based Systems*, *95*, 1–11. https://doi.org/10.1016/j.knosys.2015.11.010