2021 6th International Conference on Signal Processing, Computing and Control (ISPCC) | 978-1-6654-2554-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISPCC53510.2021.9609402

# KAS-IDS: A Machine Learning based Intrusion Detection System

Jasmeen Kaur Chahal*, Vidhyotma Gandhi, Payal Kaushal, K.R. Ramkumar, Amanpreet Kaur, Sudesh Mittal
*Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab, India*
*jasmeen.chahal@chitkara.edu.in

***Abstract*—The Internet has become an integral part of our life as we perform day-to-day work like e-Banking, e-Education and e-Commerce. With this, the threat of attackers and hackers has also been increasing. A crucial part has been played by an Intrusion Detection System (IDS) detect such malicious acts. Unfortunately, most of the commercial IDSs are based on misuse based that are developed to capture the already known attacks only. These require frequent updation of signatures and have minimum capacity to capture new attacks. Therefore, anomaly-based IDS is an effective alternative for this problem. Many of the researchers adopt various techniques to enhance the efficiency of the IDS. However, the false alarm rate and detection rate are challenging issues.**

**This paper proposes a technique called KAS-IDS, i.e. K-Means and Adaptive SVM based Intrusion Detection System. In the first step, the clusters of data have been made using K-Means and second, the classification has been performed using adaptive SVM. A well-known dataset has been used to perform the experiments. The outcomes represents that our approach has better performance as compared to the individual algorithm when it comes to detection accuracy and false alarm rate.**

***Keywords—Intrusion Detection System, Anomaly based detection, Machine Learning, Data Mining, K-Means, SVM.***

## I. INTRODUCTION

In recent years, people have become very much technology dependent. Internet usage rate has increased exponentially. There are umpteen advantages of the internet unfortunately, it comprises huge security issues. Intrusion is a term describing the malicious act of compromising the system and affects the integrity, confidentiality, and availability of the resources [1]. Even though the firewalls and routers protect the network, but they lack to detect malicious data and intruders. It is an IDS (Intrusion Detection System) that detects such malicious activities. IDS monitors the network and system activities and policy violations and produce reports to the administrator. It generates alerts when there is an occurrence of intrusion. So, IDS primarily enables computer systems to tackle the network threats[2].

In past years, various AI (Artificial Intelligence) and ML (Machine Learning) based techniques were utilized to increase the accuracy of a raw IDS. Classification and Clustering are two primarily used approaches in this field. The former is based on a supervised machine learning technique that uses predefined classes in which objects are assigned, while the latter is an unsupervised machine learning technique that identifies similarities between objects.

There are numerous classification based algorithms such as Naïve Bayes [3], Decision Tree (DT) [4], k- Nearest Neighbor (k-NN) [5], Random Forest [6], and Support Vector Machine (SVM) [7]. Out of these, SVM is highly effective and becomes the best learning algorithm for classification. This sorts the data into two series, creating a model as it is initially trained. It is successfully applied on various applications including text and hypertext classification, image processing, and biological sciences. It has recently been used in the intrusion detection fields also. It is an anomaly approach due to their excellent generalization appearance and the potential to reduce the dimensionality. Moreover, this is quite clear and consistent to choose suitable process parameters as it has minimum dependence on conventional empirical risk like neural networks. A major benefit of using this algorithm is its high speed and ability to find real-time intrusions. The scalability of SVM is highly adaptive and can expand better as the classification has no concern with the dimensionality of the feature space. Furthermore, it has the capability to add and modify the sequences automatically.

In clustering, K-means [8] is one that is extremely widely used in many applications. In this partitioning technique, a centroid has been used to form the clusters. It is a two-phase process, wherein the first phase a number of centroids (k) have been defined, i.e. one for each cluster and in the second, the collection of these has been done according to the nearest centroid. In this paper, a combination of a proposed adaptive SVM technique and K-means has been applied to identify the intrusion. NSL-KDD [9] dataset has been utilized in simulation.

Various researchers have used different anomaly and misuse-based techniques to detect the intrusions with higher accuracy and minimum delay. A comprehensive literature review is presented in Table 1.

By motivating and analyzing the literature, a hybrid approach is proposed to detect any intrusion.An appreciable number of research articles' detection technique is based on K-Means Algorithm. Although K-Means is the basic algorithm of clustering and widely used in intrusion detection system [31], this algorithm has various limitations:

- Needs a priori information regarding different clusters.
- K- Means does not able to identify the two clusters which are completely overlapping
- It starts with a random choice of cluster centres; therefore, the results may be different on different runs of an algorithm.

TABLE I.
SUMMARY OF THE PREVIOUS WORK RELATED TO THE TECHNIQUES USED FOR INTRUSION DETCTION SYSTEM

| Author/ Ref/Year | Detection Technique | Data Set | Platform | Evaluation Parameters |
|---|---|---|---|---|
| Depren and Topallar et. al [10], 2005 | Self-Organizing Map (SOM) | KDD Cup 99 | N/A | FPR-1.25%<br>DA- 99.90% |
| Feng and Zhang et al [11], 2014 | SVM with CSOACN | KDD CUP'99 | Simulation | FPR:2.776%<br>TPR:0.300 %<br>DA-78.180% |
| Kim and Lee et al. [12],2014 | decision tree algorithm and SVM | NSL-KDD data set | Weka 3.6 and Matlab | FPR:2.1%<br>DA-99% |
| Duque and Omar [13],2015 | Clustering (K-Means) | NSL-KDD | Simulation | FPR:0.74% - 31.91%<br>TPR:99.82% - 95.70%<br>DA-70.75% |
| Hu and Li et al. [14] 2015 | K-means algorithm and the FCM algorithm | DARPA 2000 | N/A | Eliminating Rate (E-R) – 95.34% |
| Ravale and Marathe et al. [15], 2015 | K-means algorithm and the SVM | KDD CUP'99 | Simulation | DA:93.33% |
| Lin and Ke et. al [16], 2015 | Cluster Center and Nearest Neighbor (CANN)<br>Classification- (k-NN Classifier) | KDD-Cup 99 | N/A | FPR-2.95%<br>DA-99.46% |
| Desale and Kumathekar et al. [17], 2015 | Naïve Bayes, hoeffding tree | NSL-KDD | Massive Online Analysis (MOA) | DA-95.00% |
| Elakar [18] 2015 | J48, Random Forest, Random Tree | KDD Cup 99 | WEKA 3.7.11 | DA-J48 – 92.617%,<br>DA-Random Forest – 92.48%<br>DA-Random Tree – 90.35% |
| Aburomman and Reaz [19], 2016 | Classification (k-NN, SVM, PSO Algorithm) | KDD99 dataset | Matlab-2012b | DA-89.02% |
| Jabbar and Aluvalu et al. [20], 2017 | K means<br>ADA (Alternating Decision Tree) tree<br>k-NN | Gure KDD cup | N/A | FPR-0.0003%<br>DA-99.8% |
| Roshan and Miche et al. [21], 2018 | CLUS-ELM | NSL-KDD dataset | N/A | FPR-0.03%<br>DA-84% |
| Yanqing and Kangfeng et al. [22], 2019 | fuzzy aggregation approach using the MDPCA and DBNs | NSL-KDD and UNSW-NB15 | Tensorflow Environment | FPR-2.62<br>DA-82.08 |
| Wang and Ouyang et al. [23], 2019 | K means<br>K-NN | KDD CUP99 | Simulator | FPR-0.29%<br>TPR-98.32%<br>DA-99.50% |
| Zhou and Cheng et al. [24], 2019 | CFS-BA<br>C4.5, Random Forest (RF) | NSL-KDD, AWID, and CIC-IDS2017 | Weka 3.8.3 | FPR-0.001%<br>DA-99.8% |
| Sandosh and Govindasamy [25], 2020 | K-NN | KDD CUP'99 | N/A | DA-92.23% |
| Nancy and Muthurajkumar et al. [26], 2020 | decision tree classification algorithm | KDD cup 1999 | N/A | DA-89% |
| Bhattacharya and Krishnan S et al. [27], 2020 | Hybrid PCA-firefly algorithm<br>XGBoost algorithm | Data Collected using Kaggle | Google Colab GPU platform | DA-99.9% |
| Mendonca et al. [28], 2021 | Tree-CNN hierarchical algorithm with the Soft-Root-Sign (SRS) | CICIDS2017 | Weka | DA-98% |
| Khan et al. [29], 2021 | Deep extreme learning machine (DELM) | NSL-KDD | Weka | DA- 91.23% |
| Zhong et al. [30], 2021 | sequential model based on Deep Learning | he KDD-99 and ADFA- | Tensorflow 2.0 | DA- 89.23% |

*DA- Detection Accuracy, FPR- False Positive Rate, TPR- True Positive Rate*

To achieve better results and accuracy, a combination of data mining techniques is required. The approach must be self-optimized, simple and fast. Instead of introducing a new clustering approach, the accuracy rate can be increased by making it hybrid with the classifier. Therefore, the proposed system is a hybrid approach of K-Means Clustering with Adaptive SVM [32] Classification which gives better accuracy. The major contribution of this paper is to work on a hybrid approach combining K-means and adaptive Support Vector Machine classifier to detect intrusions.

The next sections are ordered in this way. Section II comprises methodology and section III discusses the proposed technique. The evaluation results are discussed in part IV followed by the conclusion in part V.

## II. METHODOLOGY

This part of the paper will show the methodology used by our proposed approach. The background information of various machine learning algorithms has been given before understanding its use in the proposed methodology.

### *A. K-Means Clustering*

It is a process where the data which has similar behavior is converted into groups. Being unsupervised learning, K-means has data that is unable to specify the learning process. As shown in Table 1, there are many researchers who use this technique in many hybrid approaches to data identifying malicious data. The steps are given below:

1) Select the count of centroids identities from the dataset and make them initials ones.
2) Compute the Euclidean distance from the data identities and the centroids.
3) There is no need to change the position of the data identity that is close to the centroid. The only change is required in the case of far identities and makes them close to the nearest centroid.
4) Recompute both centroids, the earlier one and the modified one.
5) Repeat the third step until there are not stable centroids.

Mathematically, its goal is:

$$M = \sum_{a=1}^{k}\sum_{b=1}^{n} \quad d_{ab}\ (x_b\, , y_a)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

where, $d_{ab}\ (x_b,\ y_b)$ is an eculidean distance between the data point $x_b$ and the centroid $y_a$.

Euclidean distance is:

$$d(x_b\, , y_a) = \parallel\ x_b\ - y_a\ \parallel \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

### *B. Adaptive SVM*

This algorithm has the capability to use various classification techniques on different available datasets such as DARPA 2000 and NSL-KDD [32]. A thorough description of this algorithm has been given in [33].

### *C. NSL-KDD Dataset*

NSL-KDD dataset is a new version of KDD Cup 1999 dataset. This dataset has nominal count of records which makes training and testing easy. It has 41 attributes in each record and has labels (attack/normal) for each. The whole dataset is comprised of four attack classes, given below:

1. *DoS* ( Worm, Land, Smurf, and Pod)
2. *Probe* ( Portsweep, IP sweep, and Mscan)
3. *R2L (* Spy, Named, and Imap)
4. *U2R* (Xterm, Rootkit, and Perl)

### *D. Detection Approach*

Figure 1 describes the various modules of hybrid approach. The Hybrid approach of IDS is best suitable to obtain the desired detection and fake alarm rate. While K-Means algorithm has some disadvantages, it plays prominent role in this field when combined with other techniques. In our detection approach, we have combined adaptive SVM and K-Means. K-Means were used prior to classification in whole approach. This divides the train data in such a way that members belonging to one cluster take identical properties and to different clusters possess dis-similar with respect to each other.

Then, the whole data is categorized into two more datasets. The positive aspect of use of adaptive SVM is, it automatically captures the suitable variables and perform the testing with the application of Adaptive SVM dataset repeatedly on the complete dataset based on these parameters. This automatically leads to increase the detection rate. Steps for the Hybrid approach are given below:

1) A data with or without attack is loaded into NSL-KDD dataset.
2) As per the training dataset K-Mean algorithm pre-classified the data.
3) The training data is divided into two groups: normal data and anomaly data.
4) These groups are further subdivided into Train and Test data.
5) Train data is grouped into two categories: Train_train and Train_test set.

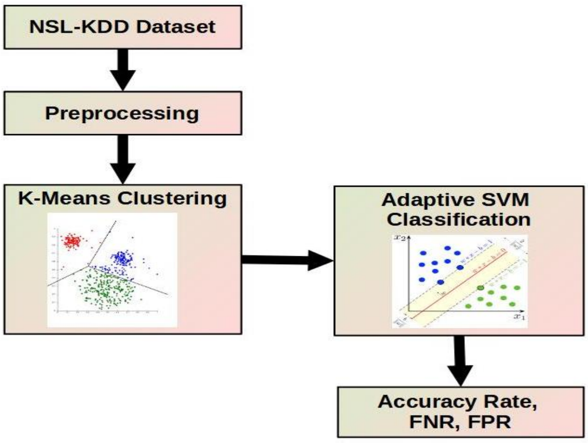

Fig. 1. Architecture of Hybrid Approach

6) Adaptive SVM Classification is applied by using some arbitrary parameters on Train_train and Train_test set.
7) Find the effective parameters.
8) Re-perform the Adaptive SVM Classification for the complete dataset.
9) Efficiency rate, False Positive rate, False Negative rate are considered as the outcomes.

.

## III. EXPERIMENTAL RESULTS

The experimental results have been evaluated by using a proposed framework as shown in Fig. 1. on NSL-KDD [9] dataset. This system requires 2.8 GHz processor. Also, in terms of RAM and storage space, it needs 2 GB and 200 GB memory respectively. Moreover, MATLAB R2013a is essential as an implementation tool. This dataset has 42 attributes in every record, and the last one can be marked either normal or attack. In order to evaluate the hybrid approach, we have compared the K- Means and Adaptive SVM as an individual approaches. The output parameters are accuracy, False Positive rate, False Negative rate.

In initial step, we have compared the performance of K-Means, adaptive SVM, and hybrid approach. As shown in Fig. 2, the proposed technique has the highest accuracy. The average detection accuracy of K-means is 81.61%, Adaptive SVM is 92.13% and the proposed hybrid approach is 99.54% as shown in Fig. 3. This means the most efficient is the combination of both.

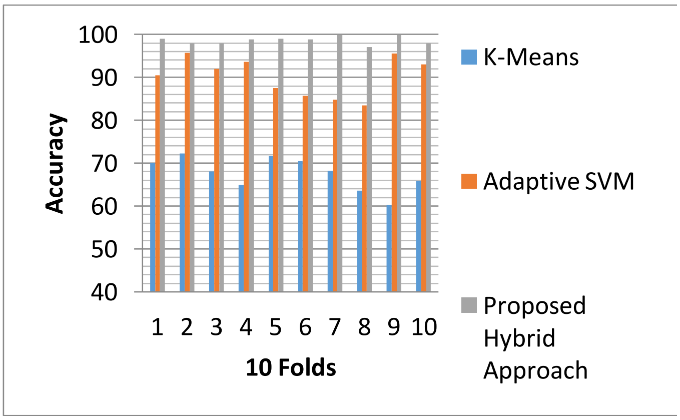

Fig. 2. Accuracy comparison of K-Means, Adaptive SVM and proposed hybrid approach

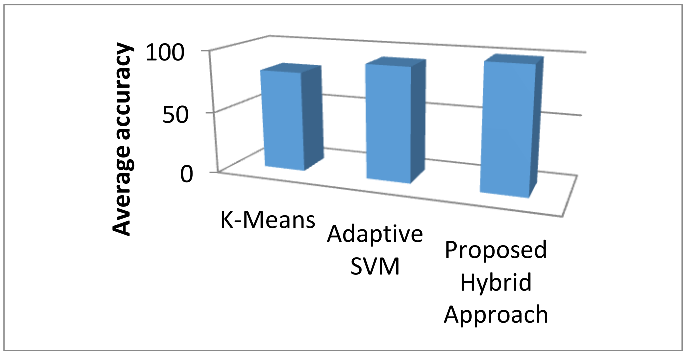

Fig. 3. Comparison of Average Accuracy Rate of three Algorithms

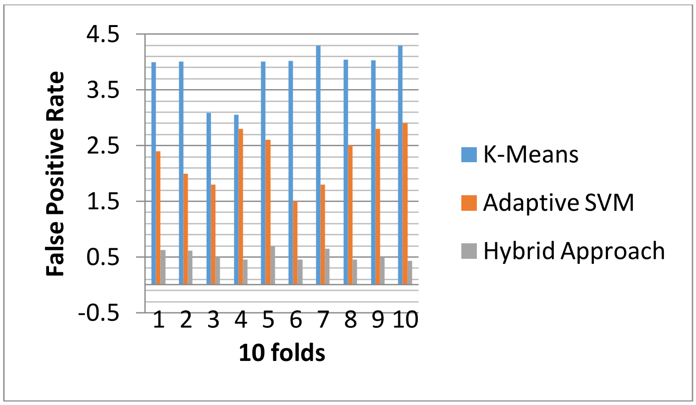

Fig. 4. Analysis of algorithms in terms of false alarms

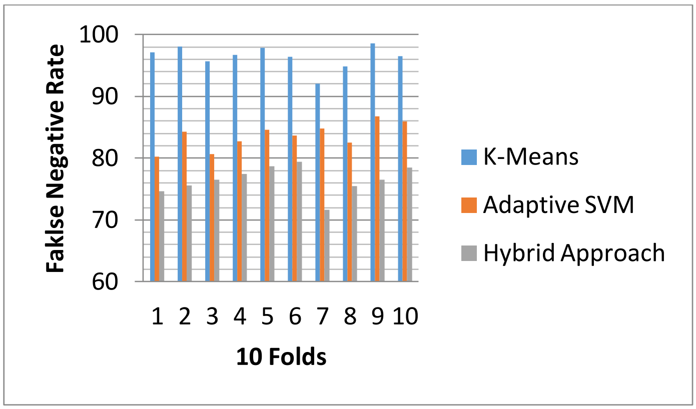

Fig. 5. Analysis of algorithms in terms of false negatives

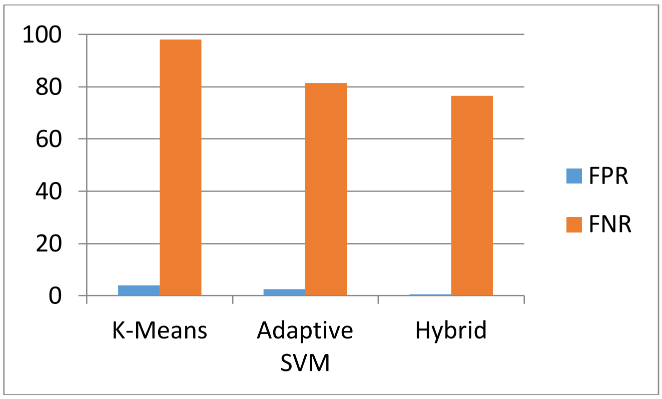

Fig. 6. Average FPR and FNR of three Algorithms

Different parameters including False Positive Rate (FPR) and False Negative Rate (FNR), and the performance of various models on such factors prove the reliability and efficiency of model. Fig. 4, 5 and 6 show the analysis and average FPR and FNR in all three algorithms where the hybrid outperforms the other.

## IV. CONCLUSION

IDS plays prominent role in identifying intrusions. This work presents a hybrid approach of K-Means and Adaptive SVM, and concludes that this amalgamation provides better results than the individual results of K-Means and Adaptive SVM. Moreover, this is highly accurate algorithm as compared to other techniques.

The hybrid technology successfully identify the data as normal and attack, and this technique is found to be

99.54% more accurate than the techniques performed solo. Therefore, it concludes that using this system in real-time gives a very high detection rate of attacks. Moreover, this approach is simple and efficient especially for the reduction of the false-positive ratio and for the rise of the false negative ratio.

## V. FUTURE SCOPE

As the present approach, the data is divided into two categories called as normal and abnormal data, and accurate results are found through NSL-KDD dataset but, it can also be utilized for the real time analysis of traffic. Apart from this, its performance in instantaneous traffic analysis can be augmented through the intelligent agents of clustering and classification algorithms.

Additionally, the experimental combination of different data mining techniques, and combinations such as artificial intelligence, soft computing and other clustering algorithms can be utilized for the improvement in the accuracy of detection. Lastly, the performance of the system can be improved by extending the intrusion detection system to intrusion prevention system.

## REFERENCES

[1] K. R. Karthikeyan and A. Indra, “ntrusion Detection Tools and Techniques – A Survey,” *Int. J. Comput. Theory Eng.*, vol. 2, no. 6, pp. 1793–8201, 2010, Accessed: Jul. 18, 2020. [Online]. [2] “Importance of Intrusion Detection System (IDS),” *Int. J. Sci. Eng. Res.* , vol. 2, no. 1, 2011, Accessed: Jul. 18, 2020. [Online].

[3] I. Rish and I. Rish, “An empirical study of the naive bayes classifier,” 2001, Accessed: Jul. 17, 2020. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.330.2788.

[4] “Decision Tree Algorithm - an overview | ScienceDirect Topics.” https://www.sciencedirect.com/topics/computer-science/decision-tree-algorithm (accessed Jul. 17, 2020).

[5] L. Peterson, “K-nearest neighbor,” *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009, doi: 10.4249/scholarpedia.1883.

[6] A. Liaw and M. Wiener, “Classification and Regression by RandomForest,” 2001. Accessed: Jul. 17, 2020. [Online]. Available: https://www.researchgate.net/publication/228451484.

[7] “1.4. Support Vector Machines — scikit-learn 0.23.1 documentation.” https://scikit-learn.org/stable/modules/svm.html (accessed Jul. 17, 2020).

[8] M. Jianliang, S. Haikun, and B. Ling, “The application on intrusion detection based on K-means cluster algorithm,” *Proc. - 2009 Int. Forum Inf. Technol. Appl. IFITA 2009*, vol. 1, pp. 150–152, 2009, doi: 10.1109/IFITA.2009.34.

[9] “NSL-KDD | Datasets | Research | Canadian Institute for Cybersecurity | UNB.” https://www.unb.ca/cic/datasets/nsl.html (accessed Jul. 19, 2020).

[10] O. Depren, M. Topallar, E. Anarim, and M. K. Ciliz, “An intelligent intrusion detection system ( IDS ) for anomaly and misuse detection in computer networks and misuse detection in computer networks,” *Expert Syst. Appl.*, vol. 29, pp. 713–722, 2005, doi: 10.1016/j.eswa.2005.05.002.

[11] W. Feng, Q. Zhang, G. Hu, and J. X. Huang, “Mining Network Data for Intrusion Detection through Combining SVM with Ant Colony,” *Futur. Gener. Comput. Syst.*, 2013, doi: 10.1016/j.future.2013.06.027.

[12] G. Kim, S. Lee, and S. Kim, “A novel hybrid intrusion detection method integrating anomaly detection with misuse detection,” *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1690–1700, 2014, doi: 10.1016/j.eswa.2013.08.066.

[13] S. Duque and M. N. Bin Omar, “Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS),” *Procedia Comput. Sci.*, vol. 61, pp. 46–51, 2015, doi: 10.1016/j.procs.2015.09.145.

[14] L. Hu, T. Li, N. Xie, and J. Hu, “False Positive Elimination in Intrusion Detection Based on Clustering,” pp. 519–523, 2015, doi: 10.1109/FSKD.2015.7381996.

[15] U. Ravale, N. Marathe, and P. Padiya, “Feature selection based hybrid anomaly intrusion detection system using K Means and RBF kernel function,” *Procedia Comput. Sci.*, vol. 45, no. C, pp. 428–435, 2015, doi: 10.1016/j.procs.2015.03.174.

[16] W. C. Lin, S. W. Ke, and C. F. Tsai, “CANN: An intrusion detection system based on combining cluster centers and nearest neighbors,” *Knowledge-Based Syst.*, vol. 78, no. 1, pp. 13–21, 2015, doi: 10.1016/j.knosys.2015.01.009.

[17] K. S. Desale, C. N. Kumathekar, and A. P. Chavan, “Efficient Intrusion Detection System Using Stream Data Mining Classification Technique,” *2015 Int. Conf. Comput. Commun. Control Autom.*, pp. 469–473, 2015, doi: 10.1109/ICCUBEA.2015.98.

[18] K. S. Elekar, “Combination of Data Mining Techniques for Intrusion Detection System,” 2015, doi: 10.1109/IC4.2015.7375727.

[19] A. A. Aburomman and M. Bin Ibne Reaz, “A novel SVM-kNN-PSO ensemble method for intrusion detection system,” *Appl. Soft Comput. J.*, vol. 38, pp. 360–372, 2016, doi: 10.1016/j.asoc.2015.10.011.

[20] M. A. Jabbar, R. Aluvalu, and S. S. S. Reddy, “Cluster Based Ensemble Classification for Intrusion Detection System,” in *ACM International Conference Proceeding Series*, 2017, pp. 253–257.

[21] S. Roshan, Y. Miche, A. Akusok, and A. Lendasse, “Adaptive and online network intrusion detection system using clustering and Extreme Learning Machines,” *J. Franklin Inst.*, vol. 355, no. 4, pp. 1752–1779, 2018, doi: 10.1016/j.jfranklin.2017.06.006.

[22] Y. Yang, K. Zheng, C. Wu, X. Niu, and Y. Yang, “Building an Effective Intrusion Detection System Using the Modified Density Peak Clustering Algorithm and Deep Belief Networks,” *Appl. Sci.*, vol. 9, no. 2, 2019, doi: 10.3390/app9020238.

[23] Q. Wang, X. Ouyang, and J. Zhan, “A Classification Algorithm Based on Data Clustering and Data Reduction for Intrusion Detection System over Big Data,” *KSII Trans. INTERNET Inf. Syst.*, vol. 13, no. 7, pp. 3714–3732, 2019.

[24] Y. Zhou, G. Cheng, S. Jiang, and M. Dai, “Building an Efficient Intrusion Detection System Based on Feature Selection and Ensemble Classifier,” *Comput. Networks*, p. 107247, 2020, doi: 10.1016/j.comnet.2020.107247.

[25] S. Sandosh, V. Govindasamy, and G. Akila, “Enhanced intrusion detection system via agent clustering and classification based on outlier detection,” *Peer-to-Peer Netw. Appl.*, vol. 13, pp. 1038–1045, 2020.

[26] P. Nancy, S. Muthurajkumar, S. Ganapathy, M. Selvi, and K. Arputharaj, “Intrusion detection using dynamic feature selection and fuzzy temporal decision tree classification for wireless sensor networks,” *IET Commun. Res.*, vol. 14, no. 5, pp. 888–895, 2020, doi: 10.1049/iet-com.2019.0172.

[27] S. Bhattacharya *et al.*, “A Novel PCA-Firefly Based XGBoost Classification Model for Intrusion Detection in Networks,” *Electronics*, vol. 9, no. 2, p. 219, 2020.

[28] R. V. Mendonca *et al.*, “Intrusion Detection System Based on Fast Hierarchical Deep Convolutional Neural Network,” *IEEE Access*, vol. 9, pp. 61024–61034, 2021, doi: 10.1109/ACCESS.2021.3074664.

[29] M. A. Khan, A. Rehman, K. M. Khan, M. A. Al Ghamdi, and S. H. Almotiri, “Enhance intrusion detection in computer networks based on deep extreme learning machine,” *Comput. Mater. Contin.*, vol. 66, no. 1, pp. 467–480, 2021, doi: 10.32604/cmc.2020.013121.

[30] M. Zhong, Y. Zhou, and G. Chen, “Sequential model based intrusion detection system for iot servers using deep learning methods,” *Sensors (Switzerland)*, vol. 21, no. 4, pp. 1–21, 2021, doi: 10.3390/s21041113.

[31] S. Duque and M. N. Bin Omar, “Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS),” in *Procedia Computer Science*, 2015, vol. 61, pp. 46–51, doi: 10.1016/j.procs.2015.09.145.

[32] J. K. Chahal and A. Amanjot Kaur, “A Hybrid Approach based on Classification and Clustering for Intrusion Detection System,” *I.J. Math. Sci. Comput.*, vol. 4, pp. 34–40, 2016, doi: 10.5815/ijmsc.2016.04.04.