# Machine Learning Techniques for Anomaly Detection in Network Traffic

Richa Singh
*Amity Institute of Information Technology*
*Amity University*
Lucknow, India
rsrinet.876@gmail.com

Nidhi Srivastava
*Amity Institute of Information Technology*
*Amity University*
Lucknow, India
nsrivastava2@lko.amity.edu.in

Ashwani Kumar
*Computer Science & Engineering*
*Sreyas Institute of Engineering and Technology*
Hyderabad, India
ashwani.kumarcse@gmail.com

2021 Sixth International Conference on Image Information Processing (ICIIP) | 978-1-6654-3361-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIIP53038.2021.9702647

***Abstract*—In today's technological era, anomaly detection is a major concern in front of network users. Due to the development of various network techniques, network users are also increased which leads to more traffic on the network, and due to this, it's very difficult to recognize the anomalous patterns. This paper discussed the overview of various ML techniques used to solve the problem of anomaly detection along with their pros and cons and also discussed here the framework/model's accuracy level. In this survey, strategies for identifying and mitigating abnormalities in network traffic are discussed and compared the result in terms of its accuracy and anomaly types. The current research gaps and important research concerns in network traffic anomaly detection are presented in detail. We hope that the analysis, comparisons, and after that, the identification of gaps will point out the researchers in the right direction for doing advanced development in this field.**

***Keywords—Network Anomaly Detection, Machine Learning Techniques, Intrusion Detection***

## I. Introduction

Digitization, Digital transformation, industry 4.0, e.tc. are the buzzwords, and the major objective is to use data and technology to increase productivity, efficiency, and accuracy. The Key enabler can extract useful information from a vast amount of data, making it possible to reduce cost, save time and optimize capacity. The extraction of useful information is done with the help of various techniques. The data analysis is performed, and data is analysed with the help of various analysis techniques. Nowadays, various network communities have faced a constant challenge about the quality of services and security in large-scale networks. Various External or internal factors can cause such security problems. Stealing security information or shut down all the services are in the category of external factors, and errors related to configuration, traffic congestion, power outages, server crashes are all the internal factors.

Apart from all these problems, one security threat, normally called an anomaly, is viral nowadays. Sometimes, the data deviated from the normal datasets, or various datasets' patterns are different from the normal dataset. This deviation is known as Anomaly that reflects a significant impact on the network services and harms the network operations. Anomaly has various definitions. According to Lakhina et al. [1], "anomalies are different patterns and slight modification in a network's traffic levels."

The overall structure of this survey paper are: Anomaly types with its applications and examples are discussed in section II. Section III discussed various approaches of ML techniques or anomaly detection and their pros and cons. Finally, section IV is discussion gives the overall highlights of this paper followed with conclusion in section V.

## II. Anomaly Detection

### A. Types of Anomaly

Based on nature, the anomaly is basically classified into three major categories[2] is shown in “fig 1”.

- Point anomalies: It means the single data instance which is different from other data instances. For Example: Detecting fraud cases of credit cards on the basis of the amount spent.
- Contextual anomalies: It refers to the abnormalities related to the specific context and suitable for time-series data. For Example, spending a lot of amount during a festive month is ok but apart from that, it is not normal.
- Collective anomalies: Collectively take the data instances and helps in detecting anomalies. For example, or some specific time interval, the rhymes of ECG are the same, and it becomes normal in the next slot.

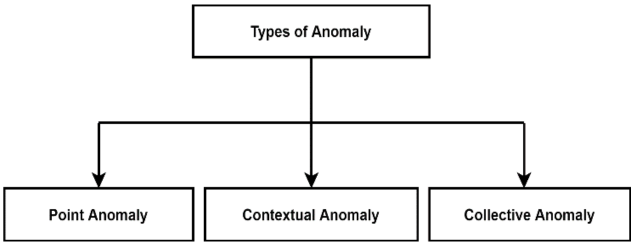

Fig 1. Types of Anomaly

### B. Machine Learning Approaches

There are various machine learning techniques used to detect an anomaly which is categorized into three: 1) Supervised, 2) Unsupervised, 3) Reinforcement Learning as shown in “fig 2”.

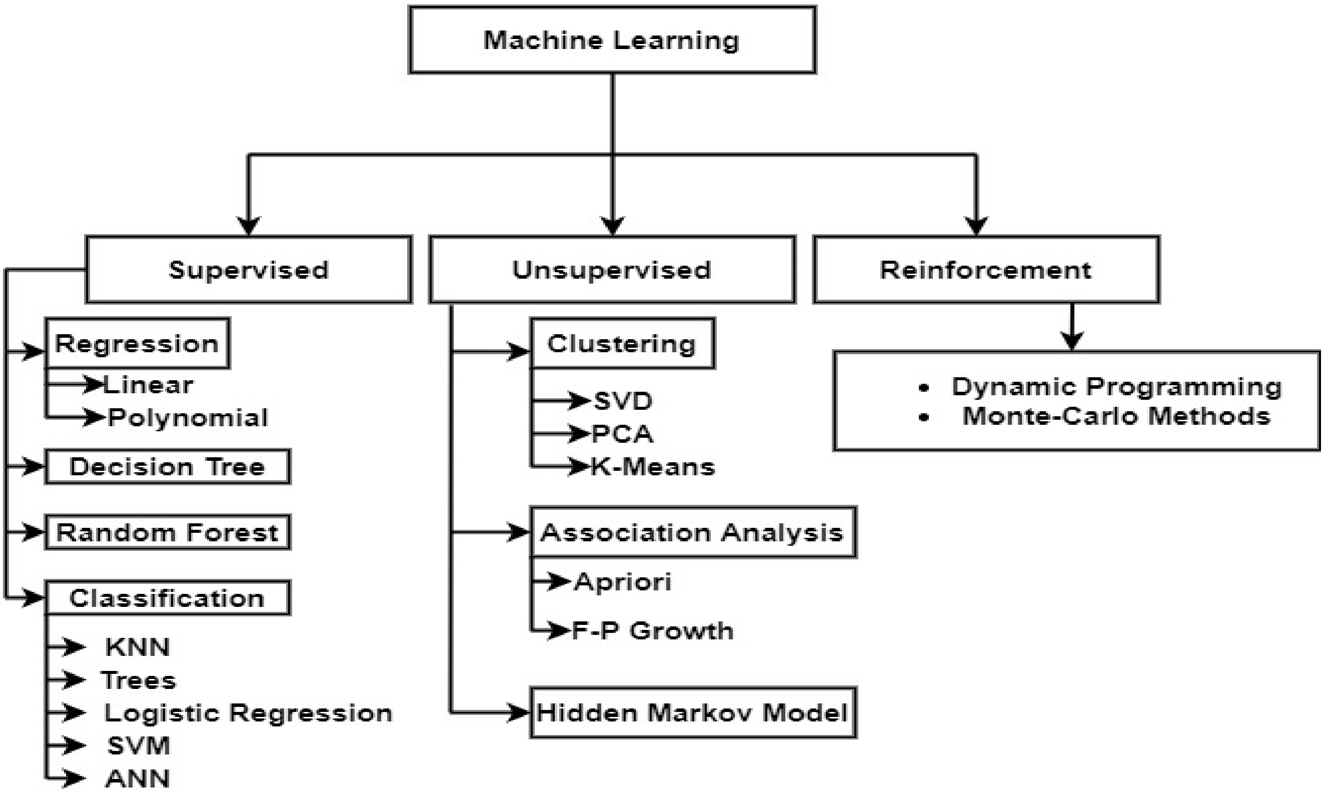

Fig2. Machine Learning Techniques

In supervised learning, various techniques are available to detect network traffic anomalies and prove their effectiveness and efficiency. Classification is also referred to as supervised learning. With the help of supervised techniques, data sets are allowed to for specific model, and instances of data are labelled using attributes sets. Supervised algorithms are the KNN, Decision Trees, Ensembles classifiers, Naïve Bayes classifier SVM and ANN. All these algorithms are applied to detect an anomaly, but every algorithm has its pros and cons. This research paper is a survey of all these techniques along with their pros and cons and accuracy level.

In an unsupervised learning algorithm, HMM and clustering are used to recognize anomalies in network traffic. İn this paper, the pros and cons of all the unsupervised algorithms are mentioned with their accuracy level.

Reinforcement learning is basically used for categorical and continuous data. So basically all are machine learning techniques and with the help of all ML techniques we can easily detect anomaly but the problem is which techniques give better accuracy. So in this paper, we discuss all the ML techniques in detail along with their pros and cons and also mention which dataset ML techniques are applied and what are their accuracy percentage. With the help of all these parameters, it's very easy to conclude which techniques give better accuracy.

### C. Motivation

One of the most significant disadvantages of network anomaly is, its nature is changing over time. So there are various ML techniques used to detect anomalies in network traffic. Now the problem is which ML techniques are better and best suited to our dataset and give us a better result. This leads to a big question mark in front of us, and this to find the the in-depth knowledge of supervised, unsupervised, and reinforcement approaches used for anomaly detection in recent applications. The major focus of literature survey is to better understand the different existing ML techniques used in Intrusion detection that may help to improve future work in this direction.

## III. Related work

In this section, a review of all the ML techniques to detect anomalies is discussed here.

Chakiret al. (2018) [3] discussed the intrusion detection model in which the effectiveness of support vector machine classifier is improved using PSO-SVM by reducing the training time and testing time. Advantages are reduction of both training as well as testing time, the effectiveness of support vector machine is improved.

Aung & Min (2018) [4] discussed the hybrid ML algorithms, i.e., k-means algorithms and Random Forest algorithms which produce good results where accuracy level is perfect for anomaly detection. This concept is valid for fixed data set, not for real-time data.

Singh R. et al.(2018)[5] This paper discussed various load balancing techniques in the grid environment along with their pros and cons.Singh R. et al.(2018)[6] This paper discussed various algorithms of load balancing techniques in distributed environments along with their pros and cons.

Weerasinghe et al. (2019) this paper [7] proposed a framework against training data integrity attacks and enhanced the resilience of support vector machines. With the help of this framework, challengers are not able to identify the specific configuration of any learner. The advantage of this framework is it helps to enhance the detection accuracy of the anomaly.

Recently, Gu et al. (2019) this paper [8] introduces a framework having the concept of Support vector machine ensemble classifier with increasing features selection. In this framework, the SVM ensemble is integrated with the powerful quality improved performance so that the training complexity is low, and performance is upgraded, higher accuracy level and low false alarm.This advantage is complexity reduced, performance upgraded, high accuracy, and high false alarm rate. The only disadvantage of this framework is that it only considered the 0/1 type of anomolus problems.

Borghesiet al. (2019) this paper [9] uses the hybrid method for recognize the anomaly, based on autoencoder. This paper shows that, the nodes learn, and on the basis of learning, it's easy to find out which one is anomalous from the normal data points. With the help of the autoencoder method, the accuracy level is increased. The advantage of this paper is to easily identify the anonymous node. Not applicable for a fixed dataset.

Chewet al. (2020) this paper [10] overcome the visibility issues of its tree rules NIDS. The pruning algorithms are modified, which is based on a decision tree. The advantage of this framework is to maintain privacy by selecting the important rules, and the second advantage is any small changes do not affect the process selection method but affect the performance of the system.

Bhatiet al.(2020) This paper [11] proposes a framework basically implemented with the help of the MATLAB tool. In this framework, an individual classifier is created and train, and based on majority voting, it takes a powerful decision. This framework has four major steps such as data collection, pre-processing, training, and testing, and the last is the decision. It provides a high detection accuracy on various datasets. The disadvantage of this paper is it provides a complex structure of the framework.

Rai (2020) In this paper [12], various ensemble learning methods are used for IDS and implemented using the python library.The advantage of this paper is that it uses the concept of a genetic algorithm and overcomes the past after-effect of DNN. The disadvantage of this paper is it is fixed with only one dataset.

Z. Ying et al.(2020) This paper [13] introduced the concept of various deep learning methods used to design a framework for anomaly detection. The advantages of this framework are better accuracy level and better adaptability. The disadvantage of this framework is using multiple deep learning methods; the structure becomes much complex. Kumar et al. proposed various object detection techniques [14-18].

W. Guanglu et al.(2020) This paper [19] introduced the concept of convolution neural networks and recurrent neural networks. This paper uses the spatial characteristics of CNN and the time series characteristic of RNN. So the comparison between these two indicates that the time series characteristics provide better performance as compared to CNN. The advantage is it provides better performance, and the disadvantage is it validates only for a fixed dataset.

P. Danel et al.(2020) This paper [20] deals with autoecoder based feature learning effects on the performance of on-network data. İn this paper, three ML techniques are used, such as autoencoder, variational autoencoder, and PCA. It provides a good result for high-dimensional data. But the disadvantage is that it uses three different ML techniques that make the system complex.

Singh R et al.(2020) This paper [21]deals with the anomaly detection framework using a convolution neural network and the accuracy percentage of anomaly.The disadvantage in this paper is that for more and complex processing the anomolus devices always leads a problem for the finding the correct detection of an anomaly.

D'Souza, D. J. et al. (2021) This paper [22] discussed the unstructured data in which outliers are detected. İt represented with the help of a graph. This paper is also a survey of static approach, dynamic approach, and ML approaches for detecting anomalies and more focused on graph-based approaches. The disadvantage is it uses multiple graph scenarios, which makes the system complex.

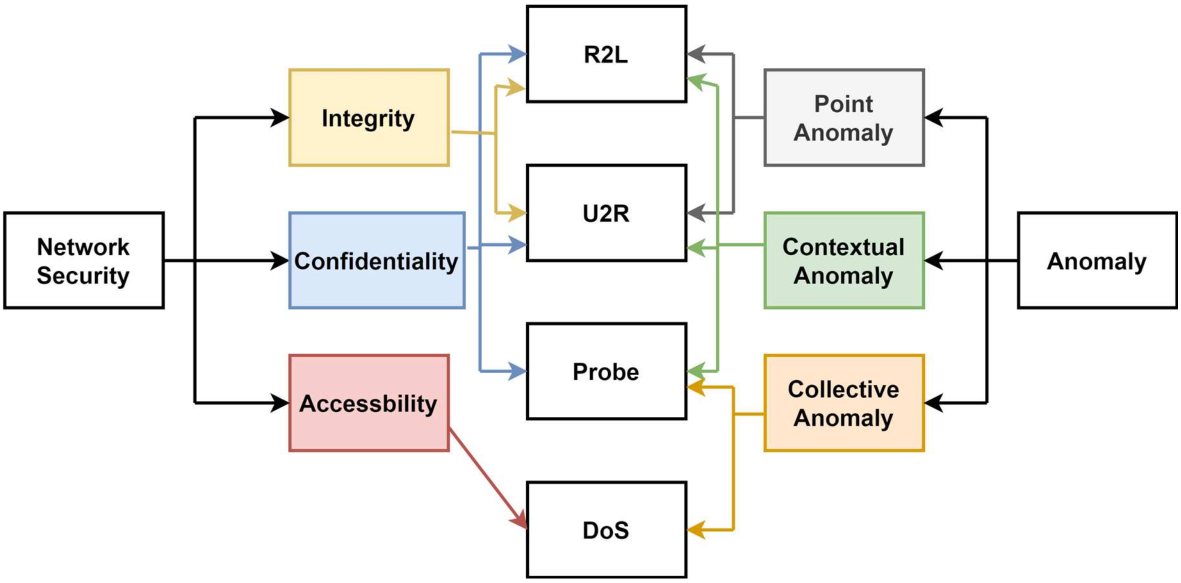

Fig 3:Relationship Between Network Anomalies and Network Attacks

The relations between network attacks and network anomalies are shown in “Fig 3”.

Table 1 shows all the research work already done that used the different supervised, unsupervised, and hybrid approach for anomaly detection. This table shows the objectives of the papers along with their advantages and challenges.

TABLE 1: ANOMALY DETECTİON APPROACHES WİTH ADVANTAGES AND CHALLENGES

| ***Authors*** | ***Year*** | ***Objective*** | ***Advantages*** | ***Challenges/Identification of Gap*** |
|---|---|---|---|---|
| Chakiret al. | 2018 | PSO and SVM are used to detect anomaly | Reduction of training and test time improved the effectiveness of the support vector machine. | This technique is used in a fixed dataset. |
| Aung & Min | 2018 | Discussed the hybrid ML algorithms, i.e., k-means algorithms and Random Forest algorithms, | Produce good results where the accuracy level is perfect. | This concept is valid for fixed data set, not for real-time data. |
| Weerasinghe et al. | 2019 | Proposed a framework to enhance the resilience of SVM, against training data integrity attacks. | It helps to increase the detection accuracy of the anomaly. | Challengers are not able to identify the specific configuration of any learner. |
| Gu et al. | 2019 | Anomaly Detection in network traffic using SVM | Accuracy level is higher | Work on a fixed dataset |
| Borghesiet al. | 2020 | Anomaly Detection in network traffic using autoencoder | Adaptability is better | It only simulates and validates a fixed dataset.<br>Accuracy % is not good |
| Chewet al. | 2020 | Overcome the visibility issues of tree rules NIDS. The pruning algorithms are modified, which is based on a decision tree. | Maintain privacy by selecting the important rules, 2. Any small changes do not affect the process selection method | Any small changes affect the performance of the system. |
| Bhatiet al. | 2020 | This paper [11] proposes a framework that is basically implemented using the MATLAB tool, an individual classifier is created and train, and based on majority voting, it takes a powerful decision. | It provides a high detection accuracy on various datasets. | It provides a complex structure of the framework. |
| Rai et. al. | 2020 | Various ensemble learning methods are used for IDS and implemented using the python library. | It uses the concept of a genetic algorithm and overcomes the past after-effect of DNN | It is fixed with only one dataset. |
| Z. Ying et al. | 2020 | A new network anomaly detection framework is introduced using multiple deep learning techniques. | Better Adaptability and accuracy in real-world environment | Complex Framework. |
| W. Guanglu et al. | 2020 | Use time series characteristics of RNN learning data and spatial characteristics of CNN learning data. The traffic model based on RNN highlights the application of deep learning technology in network security monitoring. | The performance of the RNN is better as compared to the spatial characteristics algorithm, also for network traffic anomalies it has a better detection effect. | It only simulates and validates fixed datasets. |
| P. Daniel et al. [3] | 2020 | The assessment of autoencoder based feature learning<br>the performance of a NID. | Gives better results for high dimensionality data. | For feature learning, three methods are used PCA, autoencoder and variational autoencoder make the system complex. |
| D’Souza, D. J. et. al.<br>. | 2021 | Focused on graph-based approach to detect anomaly | Adaptability is better | Often times the data does not form a network as is the case in computer networks. |

Table 2 shows all the research work already done that used the different approaches for detecting anomaly such as supervised, unsupervised, and semi-supervised. This table shows the comparison with respect to year, the algorithm used, dataset used, and accuracy percentage and anomaly types.

TABLE II: COMPARİSİON OF ANOMALY DETECTİON APPROACHES WİTH DATASET, ACCURACY LEVEL & ANOMALY TYPES

| ***Authors*** | ***Year*** | ***ML Technique/Algorithm*** | ***Dataset*** | ***Detection Accuracy (%)*** | ***Anomaly types*** |
|---|---|---|---|---|---|
| Chakir, E. M et al. | 2018 | PSO - SVM classifiers are used | NSL-KDD | 99.5% | DoS, R2L, U2R and Prob |
| Aung & Min | 2018 | k-means algorithms are used | KDD CUP | 99.9% | U2R, R2L, DoS and |
| Weerasinghe S et al. | 2019 | SVM and OCSVM are used | MNIST, CIFAR-10, SVHN | 97% | training-data integrity attacks |
| Gu et al. | 2019 | SVM are used | NSL-KDD | 99.36 % | Generic attack |
| Borghesiet al. | 2019 | Autoencoders are used | Real-time data in the network | 93.8% | generic attack in network |
| Chew et al. | 2019 | Decision trees are used | KDDcup99 | 99.33% | Generic attack |
| Bhati et al. | 2020 | Ensemble algorithm are used | KDDcup99 | 98.9 % | DoS, prob, U2R, and R2L |
| Ajeet Rai | 2020 | DNN DoS and Ensemble Methods are used | NSL-KDD | 92.7% | U2R ,R2L,DoS |
| Z. Ying et al. | 2020 | A new NAD framework is introduced using multiple deep learning techniques. | NSL-KDD | 93.2% | DoS, U2R, probing, and R2L |
| W. Guanglu et al. | 2020 | CNN and RNN are used | NSL-KDD | 95.4% | DoS, U2R, probing, and R2L |
| P. Daniel et al. | 2020 | Autoencoders are used | NSL-KDD | 94.6% | Generic attack |
| D'Souza, D. J. et. al. | 2021 | Focused on graph-based approach to detect the anomaly. | NSL-KDD | 96.5 | Generic attack |

## IV. DISCUSSION

This survey paper gives an overview of anomaly detection using various techniques. This paper gives a deep analysis of all the techniques and mentioned the advantages and future gaps, which helps the researchers to gives new concepts on the basis of the identification of gaps. This paper also gives the Comparisons in terms of accuracy level and anomaly types. There are various hybrid approaches are used, but it makes the system more complex and in the case of only a single technique, applied on the fixed dataset with better accuracy.

There is no practical implementation so far in this paper, but after analysing all the factors like dataset, anomaly type, ML techniques advantages, and identification of gap, it concludes that we can apply techniques in the real-time dataset and observe the anomaly type to find out the accuracy percentage.

## V. CONCLUSION

This paper focused on the last five year's survey on anomaly detection using the machine learning approach. This survey paper reflects all the ML techniques used for anomaly detection and also discussed the dataset, which provides better accuracy.

This paper highlights the pros and cons of all the concepts used to detect an anomaly in network traffic and on the basis of this observation we can conclude that SVM gives a better accurate result as compared to others. Besides, this survey reflects the critical challenges of anomaly detection in a real-time environment [23]. So the open issue is to detect anomalies in a real environment, not in the fixed dataset, and then calculate the accuracy level.

## ACKNOWLEDGMENT

I am highly grateful to my primary supervisor Dr. Nidhi Srivastava, co-supervisor Dr. Ashwani Kumar who guided me throughout this paper.

## REFERENCES

[1] Lakhina, A., Crovella, M. and Diot, C., "Diagnosing network-wide traffic anomalies", *In ACM SIGCOMM computer communication review,* 2004, p. 219,doi: 10.1145/ 1030194.1015492.

[2] https://blogs.oracle.com/ai-and-datascience/post/introduction-toanomaly-detection (accessed Oct. 18,2021)

[3] Chakir, E. M., Moughit, M., and Khamlichi, Y. I. "An effective intrusion detection model based on SVMwith feature selection and parameters optimization", *Journal of Theoretical and Applied Information Technology*, 2018, 96(12),pp. 3873–3885.

[4] Aung, Y. Y., and Min, M. M. "An analysis of K-means algorithm-based network intrusion detection system" *Advances in Science, Technology and Engineering Systems Journal*,3(1), 2018,pp. 496-501.

[5] Singh, R. et al. "Challenges of Load Balancing Technique in Grid Environment", *International Journal of Information Technology and Electrical Engineering*, 2018,pp. 1-5.

[6] Singh, R., & Singh, A. "Challenges of Various Load Balancing Algorithms in Distributed Environment", *International Journal of Information Technology and Electrical Engineering,* 2018,pp. 9-13.

[7] Weerasinghe, S., Erfani, S. M., Alpcan, T., &Leckie, C. "Support vector machines resilient against training data integrity attacks". *Pattern Recognition,* , 2019,pp. 96.

[8] Gu, J., Wang, L., Wang, H., & Wang, S. "A novel approach to intrusion detection using SVM ensemble with feature augmentation",*Computers & Security,* 2019, pp. 53-62.

[9] Borghesi, A., Bartolini, A., Lombardi, M., Milano, M., &Benini, L. “A semisupervised autoencoder-based approach for anomaly detection in high performance computing systems”,*Engineering Applications of Artificial Intelligence*, 2019,pp. 634-644.

[10] Chew, Y. J., Ooi, S. Y., Wong, K. S., & Pang, Y. H. “Decision Tree with Sensitive Pruning in Network-based Intrusion Detection System”. *InComputational Science and Technology*, Singapore,2020,pp.1-10.

[11] Bhati, B. S., Rai, C. S., Balamurugan, B., & Al-Turjman, F. “An intrusion detection scheme based on the ensemble of discriminant classifiers” ,*Computers & Electrical Engineering*, 2020, pp. 106742.

[12] Rai, A. “Optimizing a New Intrusion Detection System Using Ensemble Methods and Deep Neural Network”, *In2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184),* IEEE,2020,pp. 527-532.

[13] Zhong, Y., Chen, W., Wang, Z., Chen, Y., Wang, K., Li, Y., ... & Li, K. “HELAD: A novel network anomaly detection model based on heterogeneous ensemble learning”. *Computer Networks*, 2020, pp. 169.

[14] Kumar, A. and S. Srivastava, Object detection system based on convolution neural networks using single shot multi-box detector. Procedia Computer Science, 2020. **171**: p. 2610-2617.

[15] Kumar, A., Z.J. Zhang, and H. Lyu, *Object detection in real time based on improved single shot multi-box detector algorithm.* EURASIP Journal on Wireless Communications and Networking, 2020. **2020**(1): p. 1-18.

[16] Fatima, S.A., et al. Object recognition and detection in remote sensing images: a comparative study. in 2020 International Conference on Artificial Intelligence and Signal Processing (AISP). 2020. IEEE.

[17] Kumar, A., S.P. Ghrera, and V. Tyagi, Modified buyer seller watermarking protocol based on discrete wavelet transform and principal component analysis. 2015.

[18] Kumar, A., S.P. Ghrera, and V. Tyagi, An ID-based Secure and Flexible Buyer-seller Watermarking Protocol for Copyright Protection. 2017.

[19] Wei, G., & Wang, Z. “Adoption and realization of deep learning in network traffic anomaly detection device design”. *Soft Computing*, 2020, pp. 1-12.

[20] Pérez, D., Alonso, S., Morán, A., Prada, M. A., Fuertes, J. J., & Domínguez, M. “Evaluation of feature learning for anomaly detection in network traffic”. *Evolving Systems*, 2020, pp. 1-12.

[21] Singh, R., Singh, A., & Bhattacharya, P. “A Machine Learning Approach for Anomaly Detection to Secure Smart Grid Systems”. *In Advancements in Security and Privacy Initiatives for Multimedia Images,*IGI Global,2021, pp. 199-213.

[22] D’Souza, D. J., & Reddy, K. U. K. “Anomaly Detection for Big Data Using Efficient Techniques: A Review”. *Advances in Artificial Intelligence and Data Engineering,* 2021, pp.1067-1080.

[23] Cauteruccio, F., Cinelli, L., Corradini, E., Terracina, G., Ursino, D., Virgili, L., & Fortino, G.“A framework for anomaly detection and classification in Multiple IoT scenarios”. *Future Generation Computer Systems*, 2021, pp. 322-335.