IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304268 |

A Quantitative Logarithmic Transformation-Based Intrusion Detection System

Dr.S.LAVANYA, ANANTHA, ARUN KISHORE B, DINESH KUMAR C,

Professor, Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

ABSTRACT: However, the proposed model in this paper still has some shortcomings while improving the detection accuracy: Firstly, the model's ability for multi-class detection has a large room for improvement, which requires a focus on developing more powerful and scalable algorithms for processing. Secondly, it is necessary to explore the impact of incorporating more different data sources into our analysis. Thirdly, although the model has improved the detection accuracy of a small number of samples, the improvement effect is not significant. In future research, we will further study the model algorithm to further improve the accuracy of detection of a small number of samples, enhance the overall classification effect of the model, and improve the robustness of the model.

KEYWORDS: Transformer; intrusion detection; auto-encoder; NSL-KDD; deep learning

I.INTRODUCTION

The rapid growth of the internet has brought significant convenience, but it has also led to an increasing number of network security problems. In today's world, security is of paramount concern as intruders have become more sophisticated with the advancement of technology [**1**]. Hackers employ various techniques to bypass firewalls, enabling them to infiltrate network systems and cause damage to the internal infrastructure or collect individuals' private information. Given the rising threats posed by intruders, network intrusion detection has emerged as a critical research direction in network security.

Intrusion detection systems can be divided into two categories: network-based intrusion detection systems (NIDSs) and host-based intrusion detection systems (HIDSs), depending on the type of intrusion behavior being monitored [**2**]. NIDSs monitor local network traffic by examining data packets to detect intrusion behavior, while HIDSs analyze multiple sources of information collected on the local host, such as system data, log files, and disk resources. Traditional intrusion detection techniques include methods such as entropy-based approaches and redundancy optimization. Entropy-based approaches are used to detect anomalies, such as DDoS attacks in IEEE802.16-based networks, by calculating the entropy of network traffic [**3**]. This method analyzes statistical and entropy-based features of incoming traffic to determine whether an attack has occurred. However, this approach has some limitations, such as being less effective when dealing with encrypted traffic or low traffic volume. Redundancy optimization is another commonly used technique in intrusion detection, which improves the accuracy and reliability of detection by performing the same detection algorithm multiple times. The most widely used technique is Triple Modular Redundancy (TMR) [**4**], which processes input data through three detection modules and compares the output. If there is inconsistency, an attack can be detected. However, the main drawback of redundancy optimization is its high computational cost, requiring a significant amount of hardware resources and time. In recent years,

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304268 |

machine learning techniques have gained popularity in the field of intrusion detection. By analyzing large amounts of data, these techniques can discover intrusion features and patterns, leading to improved detection accuracy and speed.

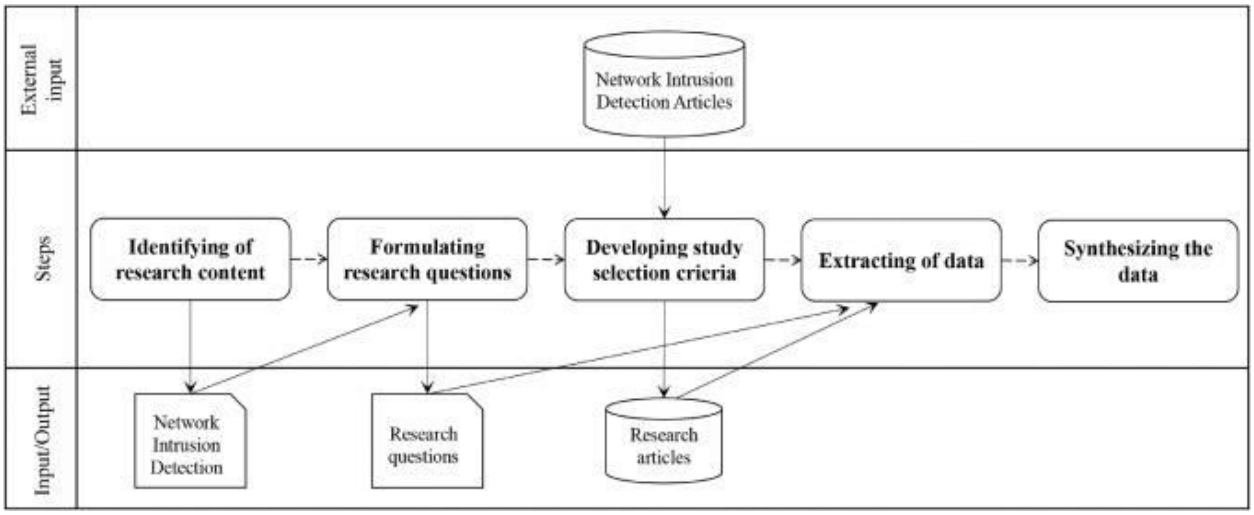

Fig 1: Transformation-Based Intrusion Detection System

As artificial intelligence continues to advance, an increasing number of algorithms have been proposed [**5**,**6**,**7**]. Researchers are now exploring the potential of applying machine learning and deep learning to network intrusion detection classification in order to improve the detection rate of intrusion detection systems.

Traditional machine learning algorithms, including decision trees [**8**], random forests [**9**], K-nearest neighbors [**10**], and Bayesian [**11**] methods, have been widely used in intrusion detection systems with good results. However, these methods may not be suitable for handling large amounts of high-dimensional data. As network technology has developed, network traffic data and features have become increasingly complex. Intrusion detection systems based on traditional machine learning algorithms may no longer be sufficient to meet the demands of network security.

II.RELATED WORK

Despite significant progress in intrusion detection using machine learning and deep learning, three key challenges remain: long model training times, imbalanced datasets, and poor performance in multi-class classification. While many researchers have proposed improvements to address these challenges, the impact of these improvements has been relatively modest.

In dealing with class imbalances, intrusion detection datasets often exhibit significant class imbalance, which may cause algorithms to favor predicting the more numerous class and thus lower detection accuracy. To address this issue, researchers often perform sampling on the dataset to balance it. suggested a detection framework that combines deep hierarchical networks with hybrid sampling techniques. In particular, they employed the one-sided selection algorithm and SMOTE technique to perform under-sampling and over-sampling, respectively, in order to balance the dataset. proposed another technique for processing unbalanced datasets, which combines SMOTE over-sampling with clustering-based under-sampling using a Gaussian mixture model. Their intrusion detection framework effectively addresses the class imbalance problem and improves detection accuracy. employed an enhanced local adaptive synthetic minority over-sampling technique

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304268 |

to address dataset imbalance and utilized an RNN for detecting various types of traffic anomalies, leading to improved accuracy in the detection process. However, these sampling methods are designed to achieve a balanced dataset in terms of quantity, without considering the issue of class overlap in the intrusion detection dataset. The difficulty of detecting data models in the overlapping regions of classes can greatly increase, resulting in a lower detection rate for the models.

In dealing with slow model training speed, the increase in feature dimensionality and quantity in intrusion detection datasets greatly extends the training time. To address this issue, researchers often utilize dimensionality reduction methods to speed up the model training process. achieved good accuracy performance by combining an auto-encoder and a residual network. They reconstructed the network using an auto-encoder to perform feature extraction, and then used the extracted features to train a designed residual network. Similarly, Liu et al. employed Principal Component Analysis (PCA) to reduce dimensionality, extracting a subset of principal component features that contain maximum information. The processed data was then fed to the recurrent neural networks for classification, resulting in a high accuracy rate. However, these dimensionality reduction methods do not take into account the loss of information caused by dimensionality reduction, which in turn results in a decrease in the model's classification ability. Moreover, slow model training speed is not necessarily only due to the increase in the number and dimensions of the dataset, as deep learning models with a deeper hierarchy and a larger number of trainable parameters can also lead to slow model training.

In dealing with the low multi-classification ability of intrusion detection models, researchers usually improve the model's multi-class detection ability through optimization. To address this issue, Hassan et al. proposed a hybrid intrusion detection model by combining convolutional neural networks (CNNs) and long short-term memory (LSTM) networks. The CNN is used to extract deep-level features, while the LSTM captures long-term dependencies between these features. To prevent overfitting, the weight matrix in the LSTM network is regularized using drop-connect. This hybrid model achieves high accuracy in intrusion detection classification. Guo et al. proposed a method for detecting attacks without any prior knowledge by combining Sub-Space Clustering (SSC) and One-Class Support Vector Machine (OCSVM). Rehman et al. proposed a combined model that employs convolutional neural networks (CNNs) and attention-based gated cyclic units for detecting both single and hybrid attacks, resulting in improved attack detection performance of the model. Yuqing et al introduced a novel method for traffic analysis by converting data traffic into pixel points in bytes. The resulting images are then processed using a CNN through operations such as convolution and pooling to obtain classification results. Their approach achieved high accuracy in both binary and multi-classification problems. However, their improvement in multi-class detection capability is not significant enough and requires further enhancement.

III.METHODS

The first part is the data processing strategy, which involves the numerical and normalization transformation of input data, dimension reduction of the dataset through the encoding layer of the stacked auto-encoder (SAE), under-sampling of normal samples using KNN, and the use of a hybrid sampling method consisting of Borderline-SMOTE for over-sampling of abnormal samples to obtain a balanced dataset, while mitigating the problem of data class overlap.

The second part is the first stage of learning, where the balanced dataset is embedded with positional information using an improved position encoding method, features are extracted using the Transformer encoder, and the binary classification model is learned and trained using softmax function.

The third part is the second stage of learning, where the binary classification model of the first stage is used to make binary predictions on the balanced dataset, the predicted values are merged with the original features to form a new dataset, which is then embedded with positional information using the improved position encoding method, features are extracted using the Transformer encoder, and the multi-class model is learned and trained using softmax.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304268 |

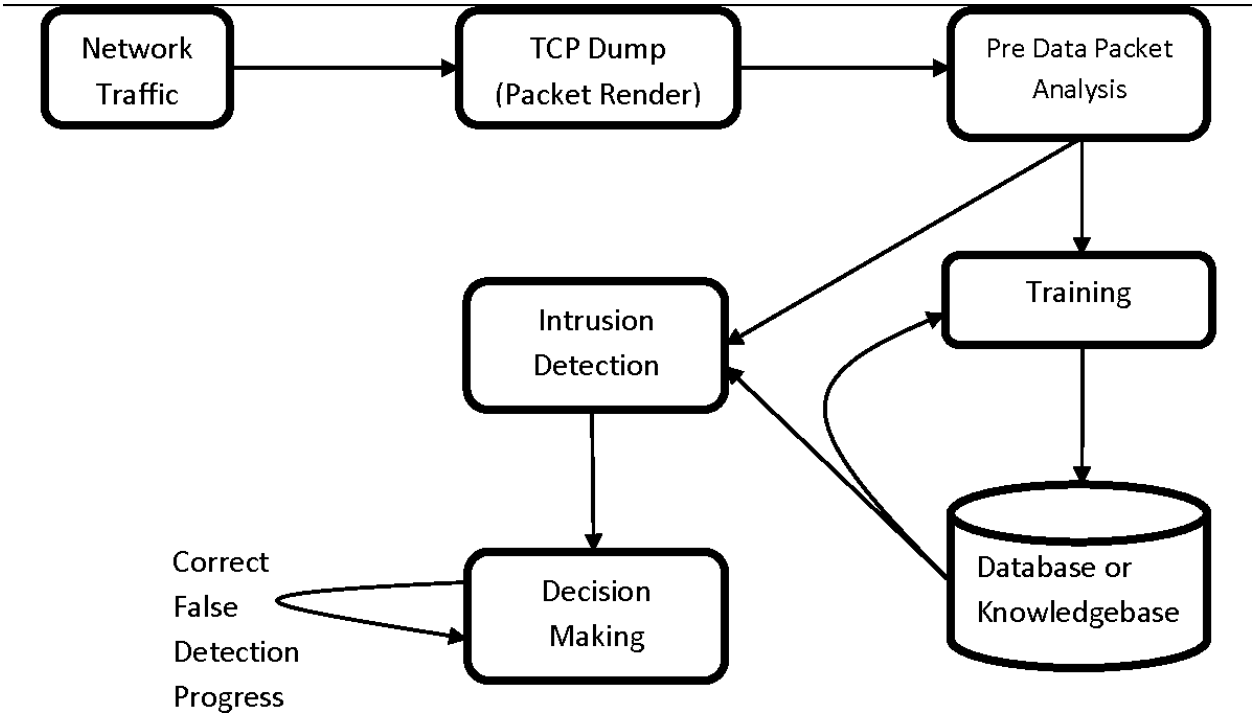

Fig 2: An overview to Software Architecture

This subsection provides a detailed explanation of the data processing approach proposed in this paper. To address the issue of high dimensionality in the dataset, which often results in long sampling and training times, we propose the utilization of stacked auto-encoders for data feature dimensionality reduction. We also introduce a hybrid sampling approach that involves under-sampling the normal samples and over-sampling the abnormal samples in the training set. In addition, we present a novel under-sampling method that leverages the properties of KNN in conjunction with the Borderline-SMOTE over-sampling method for hybrid sampling. This approach effectively resolves the challenges of class overlap and class imbalance in the dataset.

Current class imbalance problems are generally solved through under-sampling or over-sampling methods, but existing sampling methods aim to obtain a balanced dataset in terms of quantity, without considering class overlap. Intrusion detection datasets exhibit a clear class overlap problem, which can be defined as the intersection between normal and abnormal samples. In the overlapping region, even if the samples belong to different categories, their feature attributes are similar due to the similarity between feature attributes. Due to the similarity between feature attributes, the model finds it difficult to classify samples in the overlapping region, leading to a decrease in classification accuracy.

In existing under-sampling methods, random under-sampling deletes random samples, which leads to the loss of useful information even if the dataset is balanced in terms of quantity. In existing over-sampling methods, random over-sampling randomly selects samples for over-sampling, and SMOTE over-sampling randomly selects minority class samples for over-sampling. These over-sampling methods have a common feature, which is to randomly select samples for over-sampling without considering the class information of neighboring samples. This can lead to overlap between over-sampled samples and samples from different classes, making it difficult for the model to classify them. Based on the characteristics of KNN,

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304268 |

this paper proposes a new under-sampling method. Then, this paper uses this method to under-sample normal samples and over-sample abnormal samples using Borderline-SMOTE, effectively solving the problems of class imbalance and class overlap.

IV.RESULT ANALYSIS

After dividing minority class samples into three categories, the Borderline-SMOTE method can easily learn and classify Safe class samples because their neighbors are mostly from the same class. Noise class samples, on the other hand, have neighbors exclusively from the majority class and can be considered as outliers, over-sampling these samples may lead to an increase in noise and negatively affect the model performance. Danger class samples mostly have neighbors from the majority class, causing class overlap and making it difficult for the model to learn. Therefore, Borderline-SMOTE only over-samples Danger class samples to increase the number of minority samples in the overlapping region and improve the model's ability to distinguish minority samples in this area.

In intrusion detection datasets, normal samples are often more abundant than abnormal ones, which belong to the minority class. To address the issues of class imbalance and class overlap, a novel under-sampling method based on the characteristics of KNN is proposed in this paper. The method under-samples normal samples while using Borderline-SMOTE to over-sample abnormal samples. This hybrid sampling approach balances the dataset by reducing the number of normal samples and increasing the number of abnormal samples. Moreover, because the KNN under-sampling method only removes samples that belong to overlapping classes, and Borderline-SMOTE only over-samples Danger class samples, the combination of these two methods can effectively address class overlap issues.

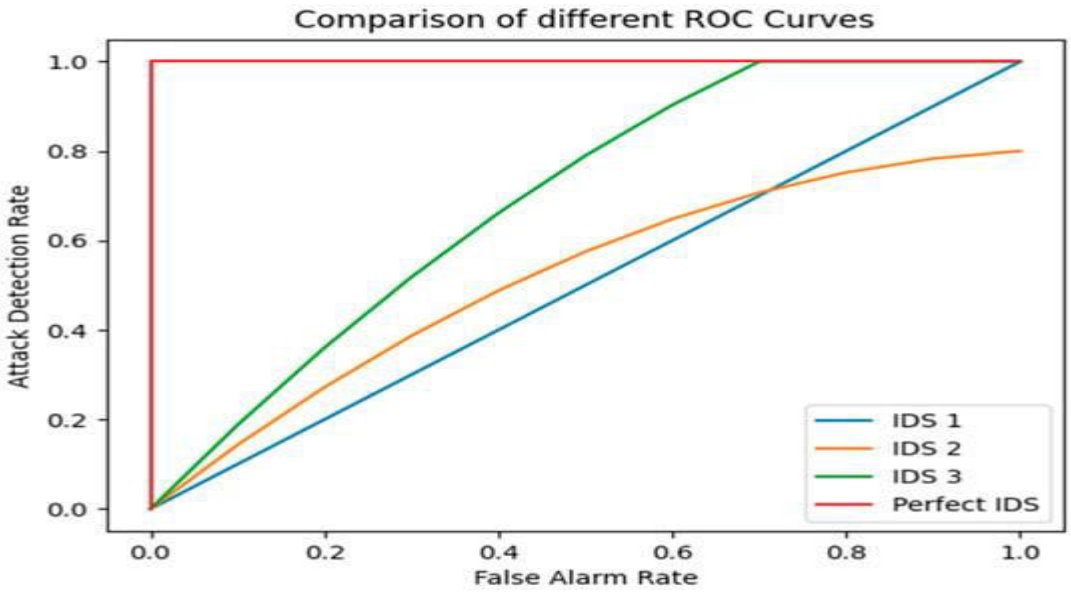

Fig 3: Result Analysis

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304268 |

It is crucial to carefully consider whether positional encoding is necessary for a particular type of data before applying it in a machine learning model. When applying positional encoding to text samples, encoding the text is a necessary step. In the positional encoding formula, the variable *pos* represents the index of each word in the text sequence, while *i* represents the index of each element in the encoded vector of the word. The positional encoding method embeds the relationship between the feature vectors of each word in the text sequence, which represents the relationship between each word. However, for intrusion detection samples, text encoding is not required, as intrusion detection samples mostly consist of numerical features. If the position is encoded according to the text samples, the variable *pos* in the positional encoding formula would represent the index of each sample, while *i* would represent the feature index of the sample. In this case, the positional encoding method would embed the relationship between each sample. However, in intrusion detection, the samples are independent of each other, and the embedded information would be irrelevant and invalid.

Although the samples in intrusion detection datasets are independent from each other, there is a correlation among their features. By embedding position codes that represent the position information of the features, the model can learn the dependency between feature positions and improve the learning performance. In other words, the model can capture the relationships between features in different positions, which can enhance its ability to learn and generalize.

V.CONCLUSIONS

This paper proposes an improved Transformer-based intrusion detection model. Specifically, this paper first proposes a data processing strategy that uses SAE to reduce the dimensionality of the dataset while ensuring the same amount of information, speeding up model training. Additionally, a KNN-based under-sampling method is proposed, combined with the Borderline-SMOTE over-sampling method to perform mixed sampling on the dataset, balancing the dataset and greatly alleviating the class overlap problem. Then, an improved position encoding method for Transformer is proposed, learning the position dependencies between features by embedding the positional relationships of features, thereby improving the model's detection ability. Finally, a two-stage learning strategy is proposed to enhance the model's multi-classification ability.

Through comparative experiments on different dimensionality reduction methods, degrees of dimensionality reduction, sampling methods, model abilities, and model ablation, this paper has demonstrated the strong detection ability and faster model training speed of the proposed model in intrusion detection models, providing promising prospects for real-time applications of intrusion detection systems.

REFERENCES

1. Wang, H.W.; Han, B.A.; Su, J.S.; Wang, X.Y. A High-Performance Intrusion Detection Method Based on Combining Supervised and Unsupervised Learning. In Proceedings of the IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Guangzhou, China, 7–11 November 2018; pp. 1803–1810. [**Google Scholar**]
2. Mishra, P.; Varadharajan, V.; Tupakula, U.; Pilli, E.S. A Detailed Investigation and Analysis of Using Machine Learning Techniques for Intrusion Detection. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 686–728. [**Google Scholar**] [**CrossRef**]
3. Shojaei, M.; Movahhedinia, N.; Tork Ladani, B. An entropy based approach for DDoS attack detection in IEEE 802.16 based networks. In Proceedings of the Advances in Information and Computer Security: 6th International Workshop, IWSEC 2011, Tokyo, Japan, 8–10 November 2011; pp. 129–143. [**Google Scholar**]

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304268 |

4. Babić, I.; Miljković, A.; Čabarkapa, M.; Nikolić, V.; Đorđević, A.; Ranđelović, M.; Ranđelović, D. Triple modular redundancy optimization for threshold determination in intrusion detection systems. *Symmetry* **2021**, *13*, 557. [**Google Scholar**] [**CrossRef**]
5. Huang, T.; Zhao, R.; Bi, L.; Zhang, D.; Lu, C. Neural embedding singular value decomposition for collaborative filtering. *IEEE Trans. Neural. Netw. Learn. Syst.* **2021**, *33*, 6021–6029. [**Google Scholar**] [**CrossRef**] [**PubMed**]
6. Zheng, J.; Lu, C.; Hao, C.; Chen, D.; Guo, D. Improving the generalization ability of deep neural networks for cross-domain visual recognition. *IEEE Trans. Cogn. Develop. Syst.* **2020**, *13*, 607–620. [**Google Scholar**] [**CrossRef**]
7. Arora, V.; Verma, K.; Leekha, R.S.; Lee, K.; Choi, C.; Gupta, T.; Bhatia, K. Transfer learning model to indicate heart health status using phonocardiogram. *Comput. Mater. Contin.* **2021**, *69*, 4151–4168. [**Google Scholar**] [**CrossRef**]
8. Ingre, B.; Yadav, A.; Soni, A.K. Decision tree based intrusion detection system for NSL-KDD dataset. In *Information and Communication Technology for Intelligent Systems (ICTIS 2017)*; Springer International Publishing: Cham, Switzerland, 2018; Volume 2, pp. 207–218. [**Google Scholar**]
9. Zhang, J.; Zulkernine, M.; Haque, A. Random-forests-based network intrusion detection systems. *IEEE Trans. Syst. Man Cybern. Syst. Part C Appl. Rev.* **2008**, *38*, 649–659. [**Google Scholar**] [**CrossRef**]
10. Liao, Y.; Vemuri, V.R. Use of k-nearest neighbor classifier for intrusion detection. *Comput. Secur.* **2002**, *21*, 439–448. [**Google Scholar**] [**CrossRef**]

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details