Performance Comparison of K-Means and t-SNE Data Compression for Intrusion Detection Using HIKARI-2021 Dataset

*Sonam Lowry[1], Mithlesh Arya[3] and Surendra Yadav[1]

[1]Department of Computer Science and Engineering, Vivekananda Global University, Jaipur, India
[2]Department of Computer Science and Engineering, JECRC University, Jaipur, India
[3]Department of Computer Science and Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur, India

sonamlowry04@gmail.com

Abstract. The problem of safety in the system is a vital as well as sensitive issue. The situation arises for the confidentiality in any institution in addition with that of the personalities too, exclusively it is significant complex data, which is conveyed diagonally to the links. In the outcome, different intrusion finding system is too much dependent on the intricate mechanism. This research paper purposes on stretch on an awareness for comparative evaluation for an unsupervised ML model using the dimensionality reduction technique. The new dataset named HIKARI-2021 is used in the projected IDS model to address the trials related to attack detection in relation to larger dimensionality, implementing innovative approaches for decrease the range, simultaneously improving the proficiency. The result presents that the comparison toward previous learning showed, taking place with the similar datasets, projected architecture achieves improved outcome for the IDS. Therefore, the final outcome is probable anticipated projected structure bids for the better rise in the trust of the security system.

Keywords: HIKARI2021, Machine Learning, Intruder Detection System.

1 Introduction

As with the growth in voluminous data attack in addition with breaches, leakages for the current time period as it may have substantial tasks of disagreement when the information is get taken by incorrect persons as of personal information may be practice, regulate in absence of authorization. [1] As though, applying safety faults, it become much added reachable in hands of invaders for authorization by apply unlawful motives. Individuals increases involved for protection, safeguarding the structures in an effect. The term intrusion defines as retrieving an ingoing in the system and precise mainframe to scout in the facts, snip information, transformation of a structure, to get safety loopholes in an operational structure for damage or terminate it. [2] Because of those effects, through this structure, use to determine interruption and reduce network

N. Vyas et al. (eds.), *Proceedings of the International Conference on Advancements in Computing Technologies and Artificial Intelligence (COMPUTATIA 2025)*, Advances in Intelligent Systems Research 189,
https://doi.org/10.2991/978-94-6463-700-7_28

and also computer risks were crucial. [3] Due to the unbalanced performance of the interruption, the structure turns up to be incapable in forecast the link stream of circulation properly, by leaks, safety necessities, like accessibility, veracity and privacy [4], so the need of emerging a progressive finding classification is improved. By this aim, research study focusses for create web safety well, overall, as emerging a precise, competent ID system. Firstly, certain data-set were pre-processed, after that system of measurement were use to decrease the count of features, intricate in construction of network classifier. Secondly, decrease the dimensionality of the detection dataset deprived of down in the significant data employing an t-SNE unsupervised model.

In remaining points of the work, it follows: In part 2 it talks about literature survey. In part 3 relays for the resources and procedures. In part 4 tells, projected prototypical and the presentation evaluations. In part 5 focuses for result and afterwords, conclusion section.

2 Literature Survey

In research of web safety, together with IDS, is constantly remained a cause for curiosity of studies everywhere in all around. The research works for the modern ID data-set labelled as HIKARI-2021was used. A limited amount of research has been shown on this dataset and will be cited in this segment. In this [5], the researcher implemented a lesser amount of availability concept on rises finding range. With usage of RF- FS technique, use of inter-packet reaching time is occurred and works as key feature. Through using this only single selection of feature, extents for data decreased.

Table 1. Survey for dataset.

Ref.	Year	Method
[5]	2023	DT, KNN, RF, MLP, Weighted KNN
[6]	2023	BiLSTM
[7]	2023	RNN, CNN, DNN
[8]	2023	Cat boost gradient boosting, Random Forest, MLP-Prod

In [6], the lightweight IOT-IDS model is used to abstract complex network material using Deep Neural Networks as that of Bidirectional Long-Short Term Memory (DNN-BiLSTM), which shows the improvement in detection of performance as well as managing the resource constraint in the dataset. In [7], the researcher proposed various deep learning classical differences to find cyberattacks in real Internet of Things (IoT) situations by means of diverse network stream pack model by retaining the strong scalar methods for data preprocessing. In [8], the study is on the identification of the unfamiliar nature of IoT links, inspects for effect combative outbreaks created information leaks in ML structures applied for cloud requests. Table 1. Shows the particulars of the associated studies cited above.

In [11], the key part of detection is, in finding and succeeding on the cyber-attacks. Whenever the combination of mining methods like, Random Tree, Random Forest, J48, Naïve Bayes are demonstrating as to be the difficult protection mechanism. In [12], answers to the tasks modeled by the progress of the cyberspace, develops the method like gaussian naïve bayes, DT and regression. In [17], it gives a complete study of the IDS. The research survey explores into a variety of numerical and ML methods. In this [18], [19] Altogether attacks are documented in this dataset stayed and approved obtainable mean IoT systems aim is to create defects in- line systems.

3 Resources with Procedures

3.1 HIKARI-2021 dataset

In [13], studied dataset HIKARI-2021 on diverse ML algorithms. With Feature Selection (FS), the data-set is concise to 22 features from that of 83. In this the balancing of the dataset is finished with that of equivalent quantity of marked classes and model reached to the accuracy of 99% and on implementation of the FS the accuracy is decreased to 80%. In [14], they anticipated a CNN with that of LSTM based IDS for locking up precision farming. The designed model is accomplished on HIKARI-2021 dataset and reached accuracy of 93.27% deprived of using any FS.

The HIKARI-2021 dataset was introduced by [9], the dataset has a whole of 555278 instances with 84 features. As of the over-all instances 347431 instances are benign and 207847 instances are attack. The facts of attack categories are given in Table 2. The HIKARI-2021data is an unstable dataset with additional benign instances. In this research, a composed dataset is prepared with 10,000 instances of benign and 10,000 instances of attack from the whole instances of 20,000. In Fig. 1, the moves were seized by means of using tcpdump with complete package capture. In case of the related traffic, by maintaining the secrecy, the anonymized collection of traffic took place. From background traffic and artificial attack, the pcap files is taken.

Table 2. Details of HIKARI-2021dataset

Network Category	Instances	Type
Background	1,70,150	Benign
Benign	3,47,432	Benign
Bruteforce	5,880	Attack
Bruteforce-XML	5,146	Attack
Probing	23,389	Attack

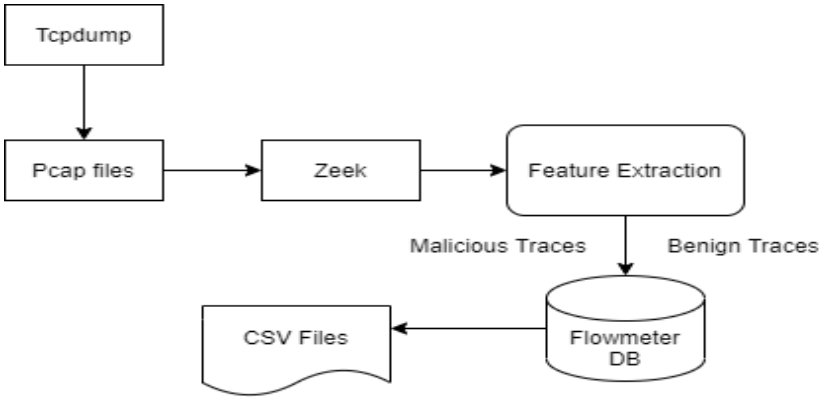

Fig. 1 HIKARI-2021 dataset preprocessing stream [9]

3.2 K-means

As concept of k-means was founded on separating of cluster. [15][16]
It calculates the remoteness between more than one information item by the help of Euclidean distance formula. By this they study, the nearer feature attributes whether lesser the remoteness amongst the items. [10] In Figure. 2 the quantity for regular points entire information set much better after the implementation of k-means as compare to the abnormal data points.

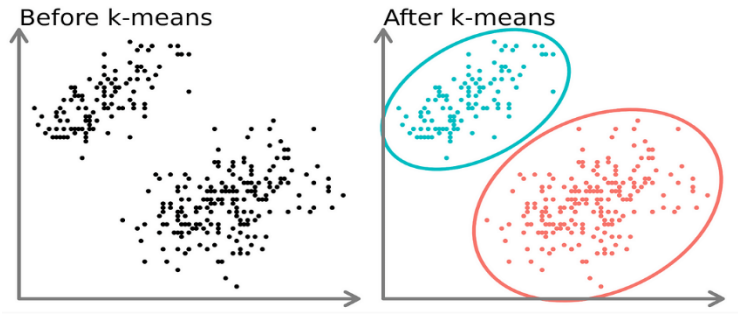

Figure. 2 Depicting data point before k-means and after k-means

3.3 t-SNE Dimensionality Reduction Technique

This particular type of unsupervised machine learning method is extensively used for falling the dimensionality of high- dimensional data. In addition, with that it efficiently protects the pairwise similarities. In Fig. 3, it illustrates the t-SNE efficacy visualization task on variable feature identification. By turning the data points into an important way for process of machine learning based data analytic structure.

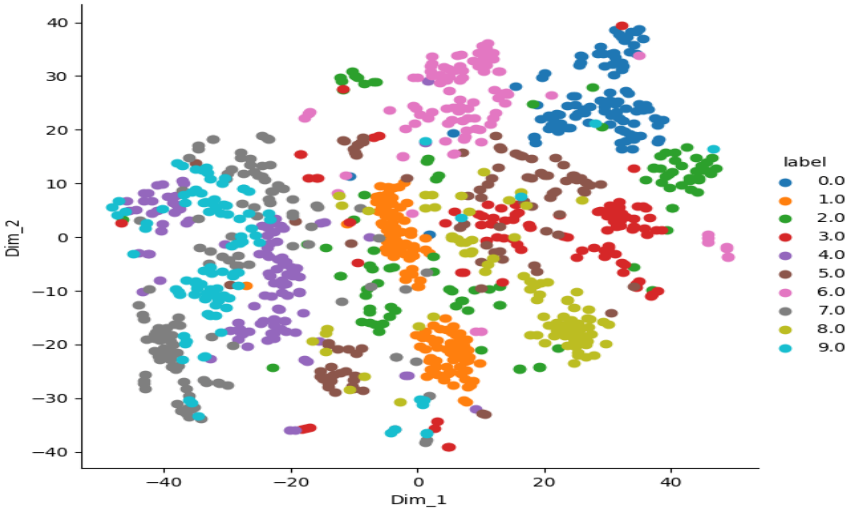

Figure. 3 t-SNE Technique

4 Proposed Model

The proposed model establishes a radical improvement in the cybersecurity, it presents a different combination of the k-means with that of t-SNE, it validated by the method which is been used for dimensionality, given Figure. 4. In this structure it includes an inclusive dataset named as HIKARI 2021 which incorporates for original and attacked traffic link forms. Succeeding with the preprocessing stage is active for the improve of the dataset, formulating it for the t-SNE application. The main motive of using t-SNE by converting complex dimensionality information in a minor way therefore on considering, tasks of implementing very well efficacy plus simplifying in a bigger explainable illustration. Therefor comparative evaluation between the outcome of the proposed model on the HIKARI-2021 dataset and former study is shown.

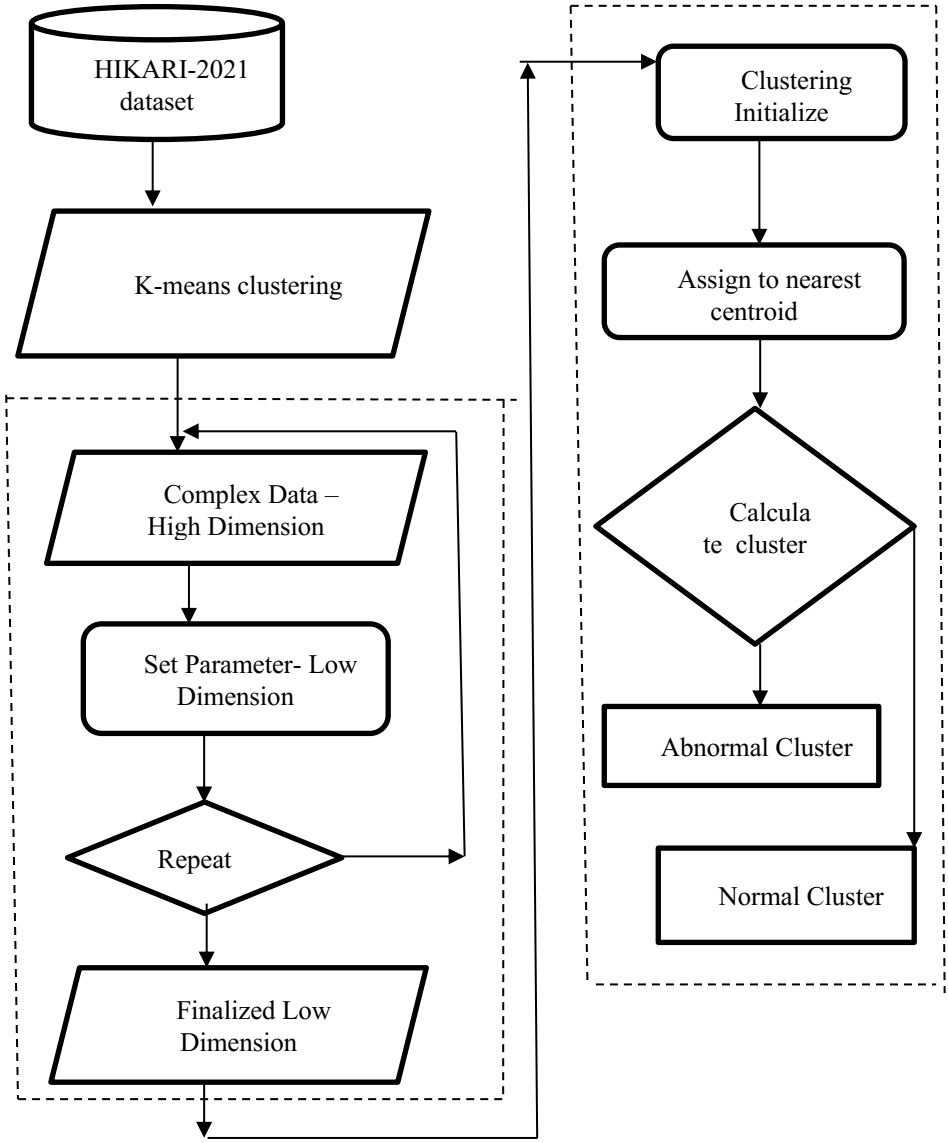

Figure. 4 Proposed IDS model by k-means + t-SNE

5 Result

The presentation of the cataloging classical model is concise and pictured using table 3. To idea an IDS with having higher accuracy and lesser testing period, two concepts are used k-means and t-SNE. The dataset was firstly, trained with k-means algorithm after that with t-SNE. The proposed prototypical model is confirmed using two key valuations system of measurement metrics like accuracy and precision in addition with testing period. In the experimenting procedure, the dataset has been divided into 80% as training whereas 20% as testing. In this, structure for MLP has attained the accuracy of 96.32% with precision of 64.26%, BiLSTM has 93.13% of accuracy with 91.80% precision, CNN has 96.52% of accuracy with 96.25% of precision, in Random Forest it has 95.25% of accuracy and precision is 95.36 whereas the proposed model gives accuracy of 97.46% and precision of 97.11%. In the resulting phase, the outcome is pictured to increase visualizations into the dataset and efficiency of the model. In Figure. 5, the t-SNE picturing shows the dimensional illustration of data, while the model results represent in a scatter plot.

Table 3. Comparison between the earlier works and proposed model

References	Accuracy%	Precision %
Ref [5]	96.32	64.26
Ref [6]	93.13	91.80
Ref [7]	96.52	96.25
Ref [8]	95.25	95.36
Proposed Model	97.46	97.11

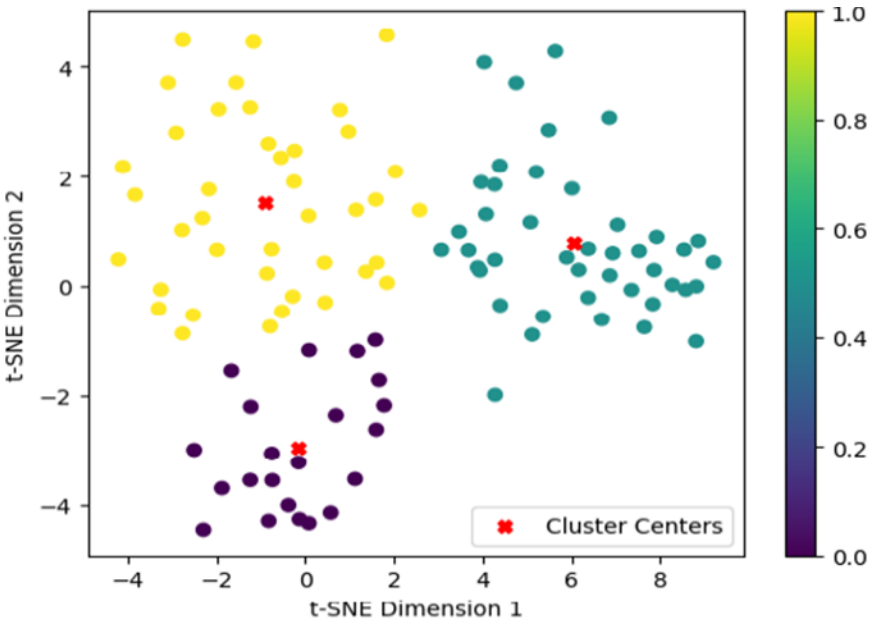

Figure. 5 k-means clustering and t-SNE visualization

Conclusion

The research study focuses in the arena of cyber security by presenting a pioneering method to increase the accuracy of the ID system. As the combination for k-means with that of t-SNE the decreasing procedure for the dimensionality, which is verifies as challenging approach. In case of old-style intrusion detection system aspects challenges in precisely finding abnormal attacked network performance in the active and multifaceted landscape of virtual threats. By using HIKARI-2021 dataset, the research modifies the k-means and t-SNE for preprocessing, follow-on with the structure that outshines in all performance system of measurement.

Declaration of competing interest

The declaration given by the authors that there will be no known competing interest regarding any parameter.

References

1. M. Abdul Quadir, J.Dibyanshu, D.Jay, H.Sabireen, I.Celetine, K.J.Sanjiv : Efficient dynamic phishing safeguard system using neural boost phishing protection, Electronics (Basel)11 3133 (2022).

2. M.Souhail, R.Tajjeeddine, A. Naseer : Network based intrusion using the UNSW-NB15 dataset, Int. J. Comput. Digit. Syst. 8 478-487, (2019).
3. H. Shoukang, L. Kai, IGAN-IDS: an imbalanced generative adversial network towards intrusion detection system in ad-hoc networks, Ad Hoc Netw. 105, 102-177, (2020)
4. K.Nilesh, T.Ritu. D. Joydip, Particle swarm optimization and feature selection for intrusion detection system, Sadhana 45, 1-14, (2022).
5. S.Onur, U.Suleyman: Advancing intrusion detection efficiency: a 'less is more' approach via feature selection, https://doi.org/10.21203/rs.3.rs-3398752/v1.(2023)
6. W.Zhendong, C.Hui, Y.Shuxin, L.Xion, L.Dahai, W.Junling: A lightweight intrusion detection method for IOT based deep learning and dynamic quantization, PeerJ Comput. Sci. 9 (2023)
7. A.Sidra, B.Imen, O.Stephen, Al H. Abdullah, A.S.Gabriel, A. Ahamad, G. Michal : Evaluating deep learning variants for cyber-attacks detection and multi-class classification in IOT networks, PeerJ Comput. Sci. (2024)
8. P.Denis, G. Lubov, Z. Artur, P.Anton, Investigation of the impact effectiveness of adversial data leakage attacks on the machine learning models, ITM Web of Conferences. 59 (2024)
9. A. Ferriyan, A. H. Thamrin, K. Takeda, J. Murai: Generating network intrusion detection dataset based on real and encrypted synthetic attack traffic. Applied Sciences, 11(17), 7868, (2021)
10. L. Wang, J. Yang, X. Xu, P. J. Wan: Mining network traffic with the K-means clustering algorithm for stepping-stone intrusion detection, Wireless Communication and Mobile Computing, vol. 2021, article 6632671, pp.1-9, (2021)
11. T. Soewu, Hemant, M. Rakhra, D. Singh: Analysis of Data Mining-Based Approach for Intrusion Detection System, In: 5th International Conference on Contemporary Computing and Informatics (IC3I), pp. 908-912 (2022)
12. Thockchom, N. Singh, M.M, U. Nandi: A novel ensemble learning- based model for network intrusion detection. Complex Intell. Syst. 9. 5693-5714 (2023)
13. R. Fernandes, N. Lopes: Network Intrusion Detection packet classification with HIKARI-2021 dataset: a study on ML algorithm, In: 2022 10th International Symposium on Digital Forensics and Security (ISDFS) pp. 1-5 (2022)
14. P. Rajak, J. Lachure, R. Doriya: CNN-LSTM- based IDS on Precision Farming for IoT data. In: IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)pp.99-103 (2022)
15. A.K. Jain, R.C. Dubes: Algorithms for clustering data, New Jersey, Prentice-Hall (1988)
16. Zhang Yufang: A kind of improved k-means algorithm, Computer Application, pp31-33 (2003)
17. S. Parhizkari: Anomaly Detection in Intrusion Detection Systems, Artificial Intelligence, Intech Open, (2024)
18. P. Adrain, P. Yuda, I.S. Redho, A. W. Eko: Improvement attack detection on internet of thinks using principal component analysis and random forest, Media J. General Comput. Sci. pp. 14-19 (2024)
19. M.S.Mousa'B, K.H. Mohammand, S. Rossilawati, I. Shayla, R.K.Atta Ur: An explainable ensemble deep learning approach for intrusion detection in industrial internet of things, IEEE Access pp. 115047-115061 (2023)