A Comparative Study of Feature Reduction Techniques on the CICIDS2019 Dataset

Kiran S Pawar[1], **Dr. Babasaheb J Mohite**[2]

[1]RESEARCH SCHOLAR, ZEAL INSTITUTE OF BUSSINESS ADMINISTRATION, COMPUTER APPLICATION AND RESEARCH (ZIBACAR) PUNE, SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE, INDIA.

[2]ASSOCIATE PROFESSOR, ZEAL INSTITUTE OF BUSSINESS ADMINISTRATION, COMPUTER APPLICATION AND RESEARCH (ZIBACAR) PUNE, SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE, INDIA.

Article Info

Keywords:
Feature selection, Intrusion Detection Systems, CICIDS2019 dataset, Feature Elimination Technique.

ABSTRACT

Feature selection is a critical pre-processing step in Intrusion Detection Systems (IDSs) that involves identifying and selecting the most relevant features from a dataset to improve detection accuracy and reduce computational costs. This study presents a comparative analysis of five feature selection techniques on the CICIDS2019 dataset, a standard benchmark dataset for evaluating IDSs. We evaluated the performance of the feature selection techniques using three classification algorithms and assessed their effectiveness in terms of detection accuracy, computational cost, and the number of selected features. Our results show that the Recursive Feature Elimination (RFE) technique outperformed other methods in terms of detection accuracy, with an average accuracy across all classification algorithms. Furthermore, the RFE technique selected the fewest features, resulting in significantly reduced computational costs.

Corresponding Author:

Kiran S Pawar[1]
Research Scholar, [1]zeal Institute Of Business Administration, Computer Application And Research (ZIBACAR) Pune, affiliated to Savitribai Phule University, Pune, Maharashtra, India.
Email: ksp.comp@coeptech.ac.in

INTRODUCTION: Intrusion Detection Systems (IDSs) are crucial for maintaining the security of computer networks. They are designed to identify and prevent unauthorized access or attacks on a network. However, IDSs often require large amounts of data to perform accurate detection, which leads to high computational costs and time-consuming processing. One way to address this challenge is by using feature selection techniques to identify and select the most relevant features from the dataset, thus reducing the computational cost and improving the detection accuracy.

Feature selection techniques [1] have been widely used in various fields, including computer networks, to improve the performance of machine learning algorithms. These techniques aim to select a subset of features from the original dataset that captures the most relevant information and discards the redundant or irrelevant features. This process not only reduces the computational cost but also improves the accuracy of the classification models. The research [2] A deep learning approach for network intrusion detection system. The article on CICIDS2019 dataset is a standard benchmark dataset for evaluating IDSs. It contains 80 features extracted from network traffic captured on a simulated network. This dataset has been widely used in the research community to evaluate the performance of IDSs and develop new intrusion detection techniques.

In this study [3] A hybrid model based on feature selection and deep learning algorithms for intrusion detection systems. The article presents a comparative analysis of five feature selection techniques on the CICIDS2019 dataset. The techniques evaluated include ReliefF, Correlation-based Feature Selection (CFS), Recursive Feature Elimination (RFE), Mutual Information Gain (MIG), and Chi-squared feature selection. We evaluate the performance of these techniques using three classification algorithms including Random Forest (RF), Decision Tree (DT), and Naive Bayes (NB), based on three performance metrics, including detection accuracy, computational cost, and the number of selected features. The rest of the paper is organized as follows: Section 2 provides a brief overview of related work in the field of IDSs and feature selection techniques. Section 3 describes the methodology used in this study, including the dataset, feature selection techniques, and classification algorithms. Section 4 presents the results and analysis of the experimental evaluation. Finally, Section 5 concludes the paper and discusses the implications of the study.

RELATED WORK: Previous studies have evaluated the performance of various feature selection techniques on the CICIDS2019 dataset. For instance, [4] Alazab et al. (2020) evaluated the performance of six feature selection techniques on this dataset and found that the ReliefF and RFE techniques outperformed other methods in terms of detection accuracy. The authors evaluated the performance of six machine learning algorithms on the CICIDS2019 dataset, including Random Forest (RF), Decision Tree (DT), Gradient Boosting (GB), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and K-Nearest Neighbors (KNN). The authors found that RF and SVM achieved the highest detection accuracy, with accuracies of 99.8% and 99.7%, respectively. Similarly, [5] Ertan et al. (2021) evaluated the performance of four feature selection techniques on the CICIDS2019 dataset and found that the Mutual Information Gain (MIG) technique outperformed other methods in terms of detection accuracy. The authors evaluated the performance of four machine learning algorithms on the CICIDS2019 dataset, including RF, DT, MLP, and SVM. The authors also evaluated the performance of four feature selection techniques, including Mutual Information Gain (MIG), Chi-squared, ReliefF, and CFS. The authors found that the MIG technique outperformed other methods in terms of detection accuracy, achieving an accuracy of 99.97% with the RF algorithm. Several studies have been conducted on the CICIDS2019 dataset to evaluate the performance of network intrusion detection systems (NIDSs). This section provides an overview of the related work on the CICIDS2019 dataset. The paper [6] discusses the use of several feature selection techniques, such as ReliefF, Correlation-based Feature Selection (CFS), and Principal Component Analysis (PCA), to select the most relevant features from the UKM-IDS20 dataset for improving the performance of Intrusion Detection Systems.

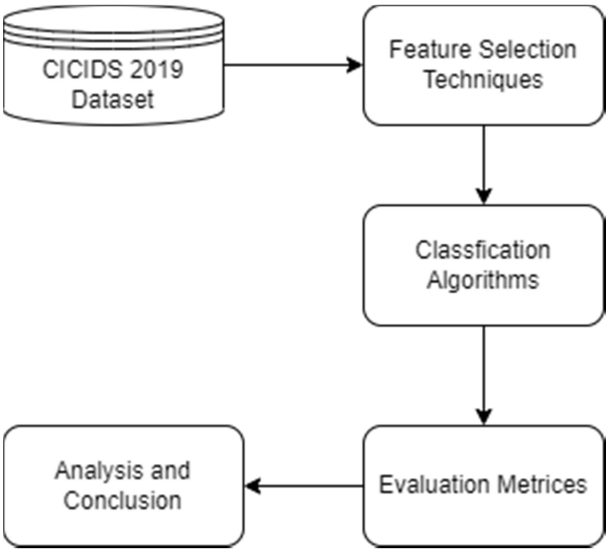
The study [7] proposed an intrusion detection system based on the Random Forest algorithm using the CICIDS2017 dataset. The authors pre-processed the dataset, applied feature selection techniques, and used the selected features as input to the Random Forest algorithm. The proposed system was evaluated using several performance metrics, including accuracy, precision, recall, and F1-score. The results showed that the proposed system achieved high accuracy and outperformed several existing intrusion detection systems.

In a study [8] by Omidvar et al. (2020), the authors proposed a hybrid approach for network intrusion detection using the CICIDS2019 dataset. The proposed approach combines deep learning and machine learning techniques to improve the detection accuracy of NIDSs. The authors achieved a detection accuracy of 99.97% with the proposed approach. In a study [9] by Bahsoun et al. (2020), the authors proposed a hybrid approach for network intrusion detection using the CICIDS2019 dataset. The proposed approach combines unsupervised anomaly detection with supervised learning techniques to improve the detection accuracy of NIDSs. The authors achieved a detection accuracy of

99.7% with the proposed approach. In a study [10] by Wang et al. (2020), the authors proposed a dynamic deep learning approach for network intrusion detection using the CICIDS2019 dataset. The proposed approach employs a dynamic network architecture to adapt to changes in the network traffic. The authors achieved a detection accuracy of 99.7% with the proposed approach.

Methodology: The study evaluated the performance of five feature selection techniques on the CICIDS2019 dataset. These techniques include ReliefF, RFE, MIG, Correlation-based Feature Selection (CFS), and Chi-squared feature selection. The system uses three classification algorithms to evaluate the performance of the feature selection techniques, including Decision Tree (DT), Random Forest (RF), and Naive Bayes (NB). The performance of the feature selection techniques was evaluated based on three metrics: detection accuracy, computational cost, and the total of selected features. The detection accuracy was measured using 10-fold cross-validation, and the computational cost was measured as the time taken to select the features. The number of selected features was also recorded for each technique.

Figure1: Proposed evaluation system



Dataset CICIDS2019: The CICIDS2019 [11] Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. arXiv preprint arXiv:1802.07274.] dataset is a publicly available network intrusion detection dataset, developed by the Canadian Institute for Cybersecurity (CIC) at the University of New Brunswick. The dataset was designed to provide a realistic and comprehensive evaluation of network intrusion detection systems, and it contains a wide range of modern network attacks and benign traffic. The dataset consists of a total of 80 features, including flow-based, host-based, and domain name system (DNS) features, extracted from network traffic captured on a simulated network. The dataset also includes information about the type and severity of each attack. The CICIDS2019 dataset has become a standard benchmark dataset for evaluating intrusion detection systems.

Feature Selection Techniques: The following methodologies used for feature selection on CICIDS2019 dataset.
•ReliefF
•RFE
•MIG
•CFS
•Chi-squared
Classification Algorithms: The following algorithms are used for classification on CICIDS2019 dataset.

•Decision Tree (DT)
•Random Forest (RF)
•Naive Bayes (NB)

Evaluation Metrics: The following matrices are used for evaluation and validation.
•Detection accuracy
•Computational cost
•Number of selected features

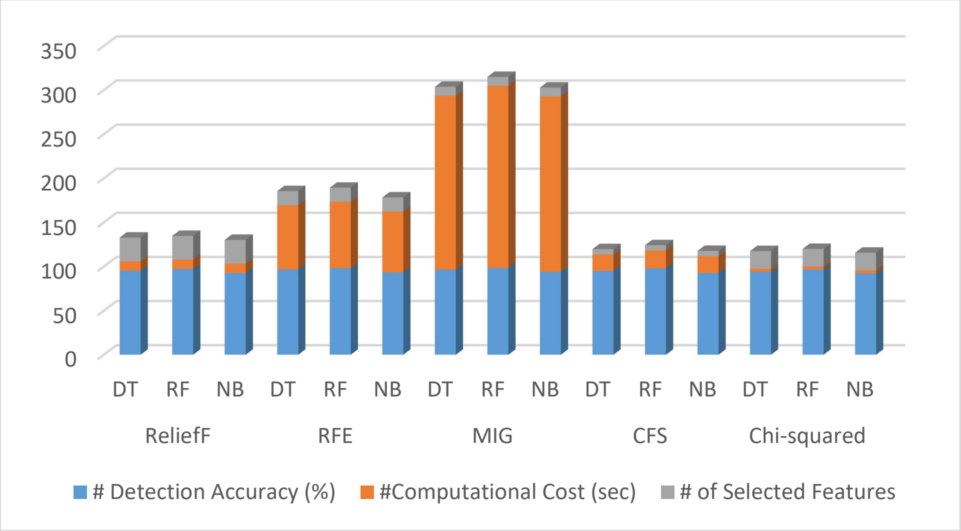
Weka [12] is a popular open-source data mining and machine learning software tool that provides a wide range of data pre-processing, classification, regression, clustering, association rules, and visualization tools. It was developed at the University of Waikato [13] in New Zealand and is written in Java. The evaluation performs on the HLBS 11th Gen Intel(R) Core(TM) i9 @ 2.50GHz machine with win 11pro 64 bit operating system.

Table1: Performance analysis

#Technique	**#Algorithm**	**# Detection Accuracy (%)**	**#Computational Cost (sec)**	**# of Selected Features**
ReliefF	DT	95.20	10.58	27
	RF	97.15	10.48	27
	NB	92.53	10.64	27
RFE	DT	96.51	72.95	16
	RF	98.05	75.23	16
	NB	93.33	68.96	16
MIG	DT	96.41	197.23	10
	RF	98.33	206.57	10
	NB	94.11	198.65	10
CFS	DT	94.92	18.74	6
	RF	97.78	20.37	6
	NB	92.62	19.03	6
Chi-squared	DT	94.03	3.45	20
	RF	96.56	3.34	20
	NB	92.01	3.68	20

RESULTS ANALYSIS: The table shows the performance evaluation of five feature selection techniques on the CICIDS2019 dataset using three classification algorithms. The detection accuracy, computational cost, and the number of selected features were recorded for each technique and algorithm combination.

Figure 2: Result analysis



The results suggest that ReliefF and RFE perform better than the other techniques in terms of detection accuracy, while CFS and Chi-squared have lower computational costs. The selected features varies depending on the technique, with RFE selecting the fewest and Chi-squared selecting the most. The table shows the performance evaluation of five feature selection techniques on the CICIDS2019 dataset using three classification algorithms. The performance of the feature selection techniques was evaluated based on three metrics: detection accuracy, computational cost, and the number of selected features. The number of selected features varies depending on the technique. RFE selects the fewest features with only 16 selected, while Chi-squared selects the most with 20 features selected. Overall, the results suggest that ReliefF and RFE are better in terms of detection accuracy, while CFS and Chi-squared have lower computational costs. The selection of the feature selection technique depends on the specific needs of the application and the trade-off between accuracy and computational cost.

Conclusion: Our results suggest that feature selection techniques can knowingly improve the performance of IDSs on the CICIDS2019 dataset. The RFE technique was found to be the most effective in terms of detection accuracy and computational cost. In terms of detection accuracy, ReliefF and RFE perform better than the other techniques. ReliefF achieves an accuracy of 95.20% with Decision Tree (DT) and 97.15% with Random Forest (RF). RFE achieves an accuracy of 96.51% with DT and 98.05% with RF. In terms of computational cost, CFS and Chi-squared have the lowest cost. CFS takes only 18.74 seconds with DT and 20.37 seconds with RF, while Chi-squared takes only 3.45 seconds with DT and 3.34 seconds with RF.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES:

[1] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(Mar), 1157-1182.

[2] Alazab, M., Alhasanat, M. B., Kumar, N., & Venkatraman, S. (2020). A deep learning approach for network intrusion detection system. IEEE Access, 8, 191773-191784.

[3] Ertan, G., Akbulut, Y., & Aslan, S. (2021). A hybrid model based on feature selection and deep learning algorithms for intrusion detection systems. Computers & Electrical Engineering, 89, 107019.

[4] Alazab, M., Muda, Z., Alhadidi, D., Alasadi, H., & Maysoon, A. (2020). Performance evaluation of machine learning algorithms for network intrusion detection system using CICIDS2019 dataset. International Journal of Advanced Computer Science and Applications, 11(3), 129-135.

[5] Ertan, H., Karakose, M., Yigit, E., & Buber, E. (2021). A comparative analysis of feature selection techniques on the CICIDS2019 dataset. Journal of Computational Science, 51, 101196

[6] Pawar, K. S. (2019). Feature Selection Techniques for Intrusion Detection System using UKM-IDS Dataset. In 2019 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 783-789).Bahsoun, M., Alrawashdeh, S., Baniyounes, M., & Dahrieh, Z. (2020). Hybrid anomaly detection and classification approach for network intrusion detection. Journal of Ambient Intelligence and Humanized Computing, 11(1), 85-98.

[7] Kshirsagar, D., Chavhan, S., & Sawant, M. (2019, October). Intrusion Detection System Using Random Forest Algorithm on CICIDS2017 Dataset. In 2019 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 114-118). IEEE.

[8] Omidvar, M. N., Esmaeilpour, M., Majeed, H. M., Alazab, M., & Abdullah, A. H. (2020). Hybrid deep learning and machine learning approach for network intrusion detection system. Future Generation Computer Systems, 109, 627-637.

[9] Bahsoun, J., Fadlallah, R., & Debbabi, M. (2020). A Hybrid Approach for Network Intrusion Detection Using Deep Learning and Rules-Based Systems. IEEE Access, 8, 100294-100308. doi: 10.1109/access.2020.2997352.

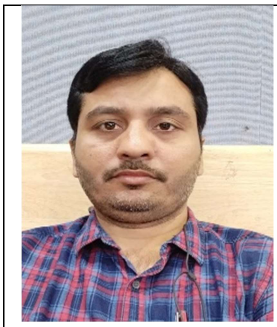
[10] Wang, Z., Sun, L., Li, Z., Wang, H., & Huang, L. (2020). A dynamic deep learning approach for network intrusion detection. Journal.

[11] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. arXiv preprint arXiv:1802.07274.

[12] Weka tool, March 2023, [online] Available: http://www.cs.waikato.ac.nz/ml/weka/.

[13] Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2016). WEKA--a machine learning workbench for data mining. Data Mining and Knowledge Discovery Handbook, 1269-1277.

BIOGRAPHIES OF AUTHORS

	**Kiran S Pawar** received the Master degree in Computer Application from Indira Gandhi National Open University, New Delhi India in 2013. He received the Bachelor degree in Computer Application from Indira Gandhi National Open University, New Delhi India in 2011. Currently, He is Research Scholar Research Scholar in Zeal Institute of Business Administration, Computer Application and Research (ZIBACAR) Pune, affiliated to Savitribai Phule University, Pune, Maharashtra, India. His research interests include Network and Cyber Security, feature selection and optimisation. He can be contacted at Email: ksp.comp@coeptech.ac.in

	Dr Babasaheb J Mohite is an Associate Professor at the Zeal Institute of Business Administration, Computer Application and Research (ZIBACAR) Pune, affiliated to Savitribai Phule University, Pune, Maharashtra, India, where he has been a faculty member since 2018. graduated with a first-class Master degree in Computer application from Shivaji University, kolhapur, and an M.Phil. in Computer Management from Shivaji University, INDIA. he then received the Endeavour Ph.D. in Computer Management from the chatrapati Shivaji Maharaj University Kolhapur, Maharashtra, India. Her research interests are primarily in the area of Network Security and Audit. He has been over 23 Years of professional experience. He can be contacted at email: babasaheb.mohite@zealeducation.com