International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 15 • Issue 1 pp. 39-46 Sept 2023-March 2024 www.csjournals.com

Enhancing IDS Performance via Feature Selection and Cyber Attack Pattern Analysis

[1]Navroop Kaur, [2]Meenakshi Bansal, [3]Sukhwinder Singh Sran
[1]Research Scholar (Ph.D.), Punjabi University, Patiala, Punjab, India.
[2]Associate Professor, CSE, Yadavindra Department of Engineering, Talwandi Sabo, India.
[3]Asst professor, Dept. of Computer Science & Engineering, Punjabi University, Patiala, India.
knavroop7488@gmail.com, ermeenu10@gmail.com, sukhwinder.sran@gmail.com

Abstract- As cyber threats continue to evolve in scale and sophistication, the role of Intrusion Detection Systems (IDS) has become increasingly crucial for protecting network environments. However, the effectiveness of IDS heavily depends on identifying and utilizing the most relevant features for accurate detection and classification. This study explores the optimization of IDS performance through targeted feature selection and advanced machine learning techniques, using two benchmark datasets: NSL-KDD and CICIDS 2018, the latter developed by the University of New Brunswick to support modern DDoS analysis. The primary objectives of this research are: (1) to identify and extract critical features corresponding to each attack category—Denial of Service (DoS), Probe, User to Root (U2R), and Remote to Local (R2L); (2) to design and implement a robust hybrid model based on eXtreme Gradient Boosting (XGBoost) to enhance detection performance; and (3) to evaluate the proposed model against established classifiers using key performance metrics. Several classification algorithms were employed, including Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT), AdaBoost (AB), Random Forest (RF), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). Experimental results demonstrated high detection accuracy across all models, with KNN, DT, and SVM achieving accuracies between 97% and 98.7%, Naïve Bayes and AdaBoost reaching up to 98.9%, and Random Forest performing exceptionally well with 99.9% accuracy. The integration of XGBoost further enhanced overall classification reliability, yielding a near-perfect accuracy of 100%.

Keywords- Intrusion Detection System (IDS), Cyber Attack Classification, XGBoost, Precision and Accuracy.

1. Introduction

In the current era of rapid digital transformation, the internet has become an essential backbone for communication, commerce, governance, and critical infrastructure. As a result, organizations and individuals rely heavily on the security and integrity of their networks and data systems. However, this increased dependency has also made networks prime targets for malicious cyber actors. The frequency, scale, and sophistication of cyber attacks are growing exponentially, ranging from simple phishing attempts to highly complex Advanced Persistent Threats (APTs) and Distributed Denial of Service (DDoS) attacks [1], [10], [15]. These threats not only compromise sensitive data but can also severely disrupt services and result in significant financial and reputational damage.

To safeguard network environments, Intrusion Detection Systems (IDS) have emerged as a fundamental component of modern cybersecurity architecture. An IDS is designed to monitor network traffic for suspicious patterns and behaviors, issuing alerts or initiating defense mechanisms when potential threats are detected [3]. IDS can be broadly categorized into signature-based detection, which relies on known patterns of attacks, and anomaly-based detection, which focuses on identifying deviations from normal behavior. While signature-based IDS are efficient at detecting known threats, they fall short when confronted with zero-day attacks or novel intrusion techniques [10], [26]. Anomaly-based systems, on the other hand, offer better detection capabilities for previously unseen attacks but often suffer from higher false positive rates [9], [19].

The effectiveness of IDS largely hinges on their ability to accurately detect and classify a wide range of cyber threats [2]. bias classifiers towards benign predictions, leading to poor detection of rare but critical attack categories such as U2R (User to Root) and R2L (Remote to Local) [2], [12].

Moreover, many IDS models are trained on outdated or unrealistic datasets that do not reflect the complexity and diversity of contemporary attack patterns. While NSL-KDD remains a widely used benchmark dataset that improves upon its predecessor KDD'99 by eliminating redundant records [6], [27], it still lacks some modern attack vectors. In contrast, the CICIDS 2018 dataset, developed by the Canadian Institute for Cybersecurity at the University of New Brunswick, provides a more comprehensive representation of current threats including botnets, brute-force attacks, and DDoS activities [24].

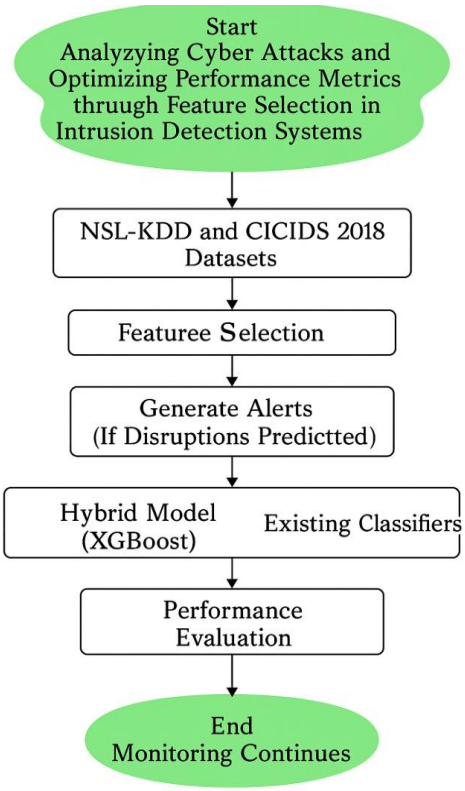

Figure 1. Real-Time Cyber Attack Detection and Response Flow in Intrusion Detection Systems (IDS)

This above Figure 1 represents a systematic approach for detecting and responding to cyber attacks using real-time data monitoring and machine learning. The process begins with continuous data collection from network activities, followed by applying feature selection techniques to extract relevant indicators for different attack types (DoS, Probe, U2R, R2L). Machine learning models like XGBoost, Decision Trees, and Naïve Bayes are used to predict potential intrusions. If threats are detected, alerts are generated, and appropriate actions are taken by the security system or administrators. This loop ensures continuous protection and improvement of the IDS.

To address the above challenges, feature selection has emerged as a critical preprocessing step in IDS development. Feature selection algorithms aim to identify the most informative and discriminative attributes in the dataset, thereby reducing dimensionality, improving model interpretability, and enhancing classification performance [4], [5], [13]. By discarding noisy or irrelevant features, IDS models can achieve faster training times and greater robustness against overfitting. In the context of cyber attack classification, selecting appropriate features for each attack category (DoS, Probe, U2R, R2L) is essential for balanced and accurate detection.

Recent advancements in ensemble learning techniques, such as eXtreme Gradient Boosting (XGBoost), have demonstrated superior performance in classification tasks involving complex and heterogeneous data [8], [11]. XGBoost is an optimized gradient-boosting algorithm that excels in speed and accuracy by employing parallel tree boosting and regularization mechanisms. Its ability to handle missing data, incorporate feature importance metrics, and reduce bias makes it an ideal candidate for building robust IDS frameworks [3], [20].

- In this research, we investigate the application of feature selection and ensemble learning methods to improve the detection accuracy and overall efficiency of IDS. Specifically, our study focuses on the following objectives:
- To extract and evaluate key features corresponding to each attack type (DoS, Probe, U2R, R2L) in the NSL-KDD and CICIDS 2018 datasets using feature selection algorithms and importance ranking.
- To design and implement a hybrid classification model using XGBoost, with the aim of enhancing precision, recall, and F1-score across diverse attack types.

To benchmark the proposed model against a suite of conventional classifiers, including Logistic Regression (LR), Naïve Bayes (NB), Decision Tree (DT), AdaBoost (AB), Random Forest (RF), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM), using performance metrics such as accuracy, precision, recall, and computational efficiency [5]. The methodology involves preprocessing two widely used benchmark datasets: NSL-KDD and CICIDS 2018. NSL-KDD is particularly useful for evaluating basic IDS concepts and

attack types, whereas CICIDS 2018 reflects realistic traffic and up-to-date attack vectors including HTTP DoS and Heartbleed exploits. We apply various feature selection methods to reduce dimensionality and isolate critical features [29]. Then, multiple classifiers are trained and evaluated under different training/testing splits. Special emphasis is placed on measuring how these models perform in detecting rare but impactful attack categories.

2.Datasets and Features Selection

In this research analyzing cyber attacks and optimizing performance metrics through feature selection in intrusion detection systems, two prominent datasets—NSL-KDD and CICIDS2018—have been employed to thoroughly evaluate the effectiveness of feature selection techniques in enhancing intrusion detection systems (IDS). The NSL-KDD dataset (https://www.kaggle.com/datasets/hassan06/nslkdd) is a refined version of the KDD'99 dataset, developed to address its major shortcomings such as duplicate entries and imbalanced class distributions. It consists of network traffic records, each described by 41 features covering protocol types, service types, connection duration, and various flow statistics. Each record is labeled as either normal or an attack type, classified under four main categories: Denial of Service (DoS), Probe, Remote to Local (R2L), and User to Root (U2R). NSL-KDD is divided into structured training and testing sets, offering a balanced and clean benchmark ideal for testing machine learning algorithms and feature selection strategies in IDS.

In contrast, the CICIDS2018 dataset https://www.kaggle.com/datasets/solarmainframe/ids-intrusion-csv) was developed by the Canadian Institute for Cybersecurity at the University of New Brunswick to provide a more realistic and comprehensive view of modern cyber threats. Unlike NSL-KDD, it captures real-world traffic collected throughout 2018, including sophisticated DDoS, Brute Force, Botnet, Port Scan, and Web attacks. The dataset comprises 80 features, encompassing both forward and backward traffic flows such as flow duration, protocol type, packet counts, and byte transfer rates. Its key field, the Label column, indicates whether a connection is normal or malicious. However, the data is inherently unbalanced, requiring careful preprocessing to prepare it for effective machine learning use.

3. Feature Selection

In this research, various feature selection techniques have been employed to enhance the performance of machine learning algorithms in intrusion detection systems (IDS) using the NSL-KDD and CICIDS2018 datasets. In the context of intrusion detection systems (IDS), applying appropriate feature selection techniques is essential for improving model accuracy, reducing overfitting, and optimizing computational efficiency—especially when dealing with high-dimensional datasets like NSL-KDD and CICIDS2018. Each dataset presents unique characteristics that demand tailored feature selection strategies. For instance, the NSL-KDD dataset, with 41 features including both categorical and numerical attributes, benefits from techniques such as LASSO Regression, Mutual Information, Chi-Square Test, and Recursive Feature Elimination (RFE). These methods are particularly effective in identifying and

preserving features that are highly relevant for classifying traditional attack types (DoS, Probe, U2R, R2L), while eliminating redundancy and noise.

Technique	Type	Description	Use in IDS	Recommended Datasets	Reason
LASSO Regression	Filter / Embedded	Adds L1 regularization to shrink irrelevant features to zero.	Useful for selecting top predictors by enforcing sparsity in high-dimensional data.	NSL-KDD, CICIDS2018	Handles sparsity; good for high-dimensional or noisy data.
PCA (Principal Component Analysis)	Dimensionality Reduction	Projects data onto lower-dimensional space capturing maximum variance.	Reduces correlated features and noise, commonly used in CICIDS2018.	CICIDS2018	Ideal for reducing 80+ continuous features to principal components.
Mutual Information (MI)	Filter	Measures the dependency between each feature and the target variable.	Identifies nonlinear relationships between features and attack types.	NSL-KDD, CICIDS2018	Versatile for both categorical and continuous features.
Correlation-Based Feature Selection (CFS)	Filter	Selects features that are highly correlated with the class and uncorrelated with each other.	Helps in removing redundant features.	CICIDS2018	Reduces multicollinearity in high-dimensional continuous data.
Chi-Square Test	Statistical Filter	Measures independence between categorical features and class labels.	Effective in NSL-KDD's categorical attributes like protocol_type, flag.	NSL-KDD	Suitable for datasets with labeled categorical variables.
Recursive Feature Elimination (RFE)	Wrapper	Recursively removes less important features using a base estimator.	Applied with SVM or RF to iteratively rank and reduce features.	NSL-KDD	Useful for feature ranking in medium-dimensional datasets.
Information Gain / Entropy	Filter	Ranks features by how much they contribute to reducing uncertainty.	Common in Decision Tree-based models and used to select splitting attributes.	NSL-KDD	Great for hierarchical classifiers like DT or RF.
Variance Threshold	Filter	Removes features with low variance as they provide little information.	Quick method to discard uninformative features.	CICIDS2018	Efficient for trimming down features with minimal variation.

Table .1 Strategic Feature Selection Techniques for Enhancing Intrusion Detection Performance in NSL-KDD and CICIDS2018 Datasets

On the other hand, the CICIDS2018 dataset contains over 80 continuous features derived from real-world traffic flows, making it more complex and suitable for modern threat scenarios like DDoS, brute-force, and botnet attacks. For this dataset, dimensionality reduction techniques such as Principal Component Analysis (PCA) and Variance Thresholding are more appropriate, as they reduce multicollinearity and focus the learning process on the most informative components. Additionally, Correlation-Based Feature Selection (CFS) is used to filter out redundant features, while LASSO ensures sparsity and model generalization.

Conclusion:

This validates the importance of combining feature selection with advanced algorithms to improve the precision and efficiency of IDS. The practical significance of this work lies in its ability to deliver fast, accurate, and scalable detection capabilities suitable for real-world deployment in varied network environments. Future work can extend the model to real-time IDS by incorporating deep learning methods like LSTM and CNN to capture temporal and dynamic threats. Addressing streaming data, concept drift, and evolving attacks will be crucial for long-term effectiveness. Additionally, federated and privacy-preserving learning can improve security and scalability in distributed systems such as IoT and cloud platforms. This research provides a strong foundation for developing intelligent and adaptive intrusion detection solutions.

References

[1] Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., & Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies, 32*(1), 1–29. https://doi.org/10.1002/ett.4150

[2] Ahsan, R., Shi, W., & Corriveau, J.-P. (2021). Network intrusion detection using machine learning approaches: Addressing data imbalance. https://doi.org/10.1049/cps2.12013

[3] Al-Imran, M., & Ripon, S. H. (2021). Network intrusion detection: An analytical assessment using deep learning and state-of-the-art machine learning models. *SN Computer Science, 1*(3). https://doi.org/10.1007/s44196-021-00047-4

[4] Alazzam, H., Sharieh, A., & Sabri, K. E. (2020). A feature selection algorithm for intrusion detection system based on pigeon inspired optimizer. *Expert Systems with Applications, 148*, 113249. https://doi.org/10.1016/j.eswa.2020.113249

[5] Almomani, O. (2020). A feature selection model for network intrusion detection system based on PSO, GWO, FFA and GA algorithms. *Neural Computing and Applications, 33*(32), 1–22.

[6] Dhanabal, L., & Shantharajah, S. P. (2015). A study on NSL-KDD dataset for intrusion detection system based on classification algorithms. *International Journal of Advanced Research in Computer and Communication Engineering, 4*(6), 446–452. https://doi.org/10.17148/IJARCCE.2015.4696

[7] Dong, R. H., Li, X. Y., Zhang, Q. Y., & Yuan, H. (2020). Network intrusion detection model based on multivariate correlation analysis - long short-time memory network. *IET Information Security, 14*(2), 166–174. https://doi.org/10.1049/iet-ifs.2019.0294

[8] Gao, J., Chai, S., Zhang, B., & Xia, Y. (2019). Research on network intrusion detection based on incremental extreme learning machine and adaptive principal component analysis. *Energies, 12*(7), 1223. https://doi.org/10.3390/en12071223

[9] Gao, X., Shan, C., Hu, C., Niu, Z., & Liu, Z. (2019). An adaptive ensemble machine learning model for intrusion detection. *IEEE Access, 7*, 82512–82521. https://doi.org/10.1109/ACCESS.2019.2923640

[10] García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security, 28*(1–2), 18–28. https://doi.org/10.1016/j.cose.2008.08.003

[11] Gu, J., Wang, L., Wang, H., & Wang, S. (2019). A novel approach to intrusion detection using SVM ensemble with feature augmentation. *Computers & Security, 86*, 53–62. https://doi.org/10.1016/j.cose.2019.05.022

[12] Karatas, G., Demir, O., & Sahingoz, O. K. (2020). Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset. *IEEE Access, 8*, 32150–32162. https://doi.org/10.1109/ACCESS.2020.2973219

[13] Kasongo, S. M., & Sun, Y. (2020). Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset. *Journal of Big Data, 7*(1). https://doi.org/10.1186/s40537-020-00379-6

[14] Kasongo, S. M., & Sun, Y. (2020). A deep long short-term memory based classifier for wireless intrusion detection system. *ICT Express, 6*(2), 98–103. https://doi.org/10.1016/j.icte.2019.08.004

[15] Kayode Saheed, Y., Idris Abiodun, A., Misra, S., Kristiansen Holone, M., & Colomo-Palacios, R. (2022). A machine learning-based intrusion detection for detecting internet of things network attacks. *Alexandria Engineering Journal, 61*(12), 9395–9409. https://doi.org/10.1016/j.aej.2022.02.063

[16] Khan, N., C, N., Negi, A., & Thaseen, S. (2020). Analysis on improving the performance of machine learning models using feature selection technique. In *Proceedings* (pp. 69–77). https://doi.org/10.1007/978-3-030-16660-1_7

[17] Kumar, V., Sinha, D., Das, A. K., Pandey, S. C., & Goswami, R. T. (2020). An integrated rule based intrusion detection system: Analysis on UNSW-NB15 data set and the real time online dataset. *Cluster Computing, 23*(2), 1397–1418. https://doi.org/10.1007/s10586-019-03008-x

[18] Kwon, D., Kim, H., Kim, J., Suh, S. C., Kim, I., & Kim, K. J. (2019). A survey of deep learning-based network anomaly detection. *Cluster Computing, 22*, 949–961. https://doi.org/10.1007/s10586-017-1117-8

[19] Latah, M., & Toker, L. (2018). Towards an efficient anomaly-based intrusion detection for software-defined networks. *IET Networks, 7*(6), 453–459. https://doi.org/10.1049/iet-net.2018.5080

[20] Naseer, S., Saleem, Y., Khalid, S., Bashir, M. K., Han, J., Iqbal, M. M., & Han, K. (2018). Enhanced network anomaly detection based on deep neural networks. *IEEE Access, 6*(8), 48231–48246. https://doi.org/10.1109/ACCESS.2018.2863036

[21] Rathore, S., & Park, J. H. (2018). Semi-supervised learning based distributed attack detection framework for IoT. *Applied Soft Computing Journal*, *72*, 79–89. https://doi.org/10.1016/j.asoc.2018.05.049

[22] Roshan, S., Miche, Y., Akusok, A., & Lendasse, A. (2018). Adaptive and online network intrusion detection system using clustering and extreme learning machines. *Journal of the Franklin Institute*, *355*(4), 1752–1779. https://doi.org/10.1016/j.jfranklin.2017.06.006

[23] Saad Alqahtani, A. (2021). FSO-LSTM IDS: Hybrid optimized and ensembled deep-learning network-based intrusion detection system for smart networks. *The Journal of Supercomputing*, *78*, 9438–9455. https://doi.org/10.1007/s11227-021-04285-3

[24] Sharafaldin, I., Gharib, A., Lashkari, A. H., & Ghorbani, A. A. (2017). Towards a reliable intrusion detection benchmark dataset. *Software Networking*, *2017*(1), 177–200. https://doi.org/10.13052/jsn2445-9739.2017.009