Research Journal of Nanoscience and Engineering
Volume 5, Issue 1, 2021, PP 15-19
ISSN 2637-5591
DOI: https://doi.org/10.22259/2631-5591.0501003

A Hybrid Approach for Feature Selection Analysis on the Intrusion Detection System Using Naive Bayes and Improved BAT Algorithm

Naga Charan Nandigama

***Corresponding Author:** Naga Charan Nandigama. Email: nagacharan.nandigama@gmail.com*

ABSTRACT

In the last two decades, global internet usage has expanded to over 4 billion users, leading to a concurrent exponential rise in malicious network activities. To safeguard networks, Anomaly Detection Systems (ADS) are deployed; however, their efficiency is frequently hampered by high-dimensional data containing duplicate or irrelevant features. This paper proposes a novel feature selection method, the Naive Bayes Improved BAT Algorithm (NB-IBA), which utilizes entropy-based concepts to identify optimal feature subsets. The proposed method is evaluated using the CICIDS2017 dataset across multiple classifiers, including J48, Random Forest, Random Tree, and Bayesian Networks. Experimental results indicate that feature reduction significantly enhances performance, with the Random Forest classifier achieving superior accuracy across various attack vectors

Keywords: *Intrusion Detection System (IDS), Feature Selection, Swarm Intelligence, BAT Algorithm, Naive Bayes, Machine Learning.*

INTRODUCTION

Over the last two decades, internet adoption has grown to approximately 4 billion users worldwide. This exponential growth has inadvertently led to a surge in malicious activities and cyber threats. To counter these threats, Anomaly Detection systems are employed to monitor network traffic. However, a significant challenge in current Intrusion Detection Systems (IDS) is the presence of duplicate or irrelevant features in network traffic data, which reduces detection accuracy and increases computational overhead.

This paper addresses this challenge by proposing an optimal feature selection method. The study utilizes a hybrid approach combining a Naive Bayes classifier with an Improved BAT Algorithm (NB-IBA) based on entropy concepts to select "disguising" (most relevant) features. By confining the study to anomaly-based IDS, we aim to enhance the performance of classification algorithms such as J48, Random Forest, and Random Tree.

Classification Algorithms

Classification is a supervised learning technique used to categorize data into predefined classes.

- Decision Trees: These structures represent dataset properties via internal nodes and decision rules via branches. Algorithms like C4.5 and CART utilize Information Gain and Gini Impurity to determine optimal splits.
- K-Nearest Neighbors (KNN): A lazy learner that classifies new instances based on similarity (Euclidean distance) to existing data.
- Naive Bayes: Based on Bayes' theorem, this probabilistic classifier assumes independence between features. It is particularly effective for text classification and high-dimensional data.

Swarm Intelligence

Swarm Intelligence (SI) mimics the collective behavior of decentralized, self-organized systems in nature, such as ant colonies or bird flocks.

- BAT Algorithm (BA): Inspired by the echolocation behavior of microbats, BA utilizes frequency, velocity, and position updates to find optimal solutions in continuous and discrete spaces.

PROPOSED METHODOLOGY

The proposed work follows a four-step process to enhance anomaly detection:

- Pre-processing: Removal of noise and handling of missing data.
- Feature Selection (NB-IBA): The Naive Bayes and Improved BAT Algorithm are used to extract optimal features from the master dataset.
- Model Training: Models are trained using 10-fold cross-validation.
- Classification: The reduced feature set is fed into classifiers (J48, Random Forest, Random Tree, Bayesian Network).

Mathematical Model of the BAT Algorithm

The movement of the search agents (bats) is defined by updating their frequency, velocity, and position.

The frequency for bat is calculated as:

$$f_i = f_{min} + (f_{min} - f_{max})\beta$$

Where in $\beta \in [0,1]$ is a random vector.

The velocity at time step t is updated based on the best global position

$$v_i^t = v_i^{t-1} + \left[x_i^{t-1} - x_*\right]f_i$$

The new position is updated as:

$$x_i^t = x_i^{t-1} + \in A^t$$

To enhance local search, a random walk is applied if a random number >

$$x_{new} = x_{old} + \in A^t$$

Where is the average loudness and $\in \in [-1,1]$.

As the bats approach a target (optimal solution), loudness decreases and pulse rate increases:

$$A_i^{t-1} = \alpha A_i^t$$

$$r_i^{t+1} = r_i[1 - \exp(-\gamma t]$$

Where α and are constants.

EXPERIMENTAL SETUP

Dataset Description

The study utilizes the CICIDS2017 dataset, introduced by the Canadian Institute for Cybersecurity. It contains 79 features and covers benign traffic and common attacks.

- Training Set: 80% (184,074 records).
- Testing Set: 20% (46,018 records).

Attack Classes

The dataset is categorized into the following attack types for this study:

- Attack-1: DoS/DDoS
- Attack-2: Port Scan
- Attack-3: Bot
- Attack-4: Web Attack
- Attack-5: Infiltration
- Attack-6: Brute Force

RESULTS AND ANALYSIS

The experimental analysis compares the accuracy of four classifiers (Random Forest, Bayesian Network, Random Tree, J48) across three different feature set sizes (n=15, n=28, and n=35).

Data Analysis (Reconstructed from Experimental Results)

The following data represents the accuracy metrics derived from the experimental analysis

ATTACK TYPE	RANDOM FOREST (RF)	BAYESIAN NETWORK (BN)	RANDOM TREE (RT)	J48
Normal	0.951	0.934	0.951	0.952
Attack-1 (DoS)	0.989	0.961	0.989	0.989
Attack-2 (Port Scan)	0.987	0.985	0.987	0.987
Attack-3 (Bot)	0.702	0.975	0.721	0.709
Attack-4 (Web)	0.124	0.983	0.124	0.118
Attack-5 (Infiltration)	0.452	0.364	0.598	0.373
Attack-6 (Brute Force)	0.985	0.986	0.985	0.986

Performance Analysis

- Feature Set Optimization: The study evaluated feature sets of size 15, 28, and 35. Optimal results were generally observed with n=28 and n=35, indicating that reducing features too aggressively (n=15) may result in information loss for complex attacks like Infiltration.
- Classifier Comparison
  - Random Forest: Consistently outperformed other algorithms in detecting DoS (Attack-1), Port Scans (Attack-2), and Brute Force (Attack-6) attacks with accuracy exceeding 98%.
  - J48 & Random Tree: Showed high accuracy for DoS and Port Scans but struggled significantly with Web Attacks (Attack-4), yielding accuracy as low as 7.1% to 12.4%.
  - Bayesian Network: While generally lower in accuracy for "Normal" traffic compared to Random Forest, it showed surprisingly higher resilience in detecting Bot attacks (Attack-3) and Web attacks (Attack-4) in the n=28 configuration.

The analysis reveals that Random Forest consistently performs as a top-tier classifier across most attack categories.

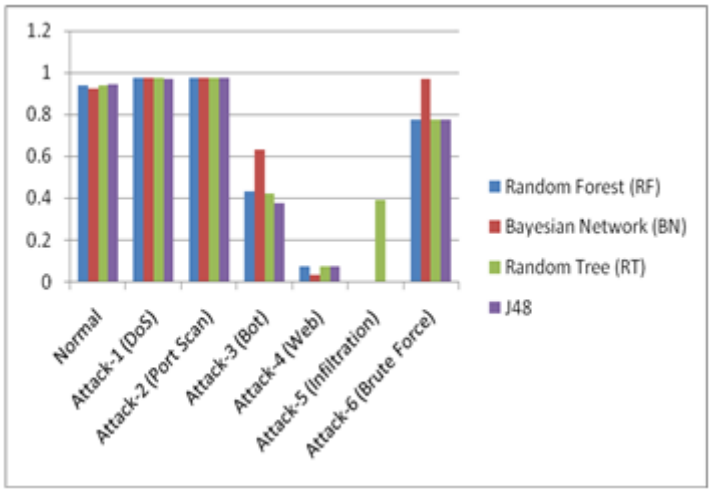

Fig1. *Feature Set n=15: The classifiers struggled with complex attacks (Web and Infiltration). Attack-5 (Infiltration) was largely undetected by RF, BN, and J48.*

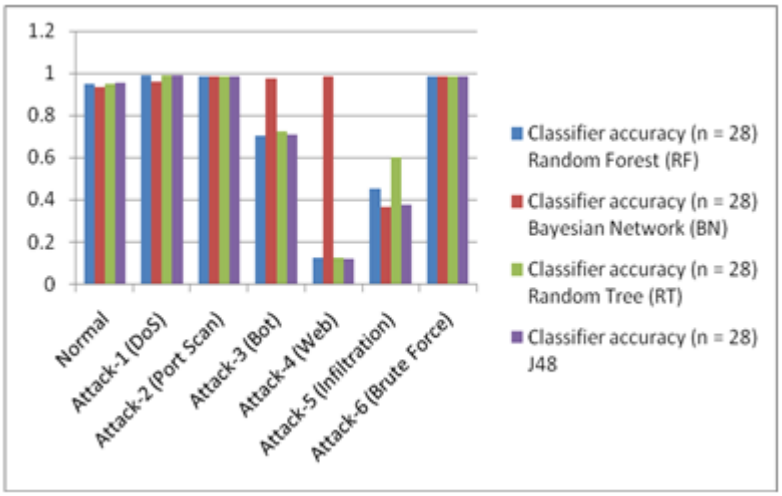

Fig2. *Feature Set n=28: There was a marked improvement in detection. Random Forest, Random Tree, and J48 achieved high accuracy in detecting Attack-1 (DoS), Attack-2 (Port Scan), and Attack-6 (Brute Force). However, Bayesian Network proved superior for Attack-3 (Bot) and Attack-4 (Web) in this configuration.*

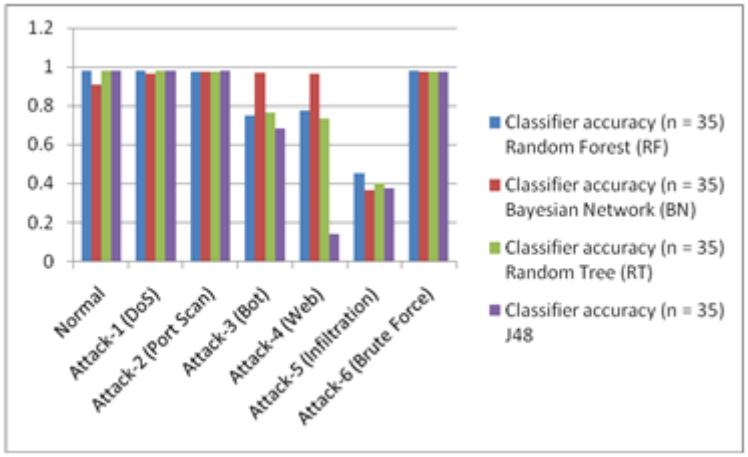

Fig3. *Feature Set n=35: Increasing the feature set to 35 provided the most stable results for Random Forest, particularly improving the detection of Web Attacks (0.774) compared to smaller feature sets.*

Summary of Findings

Among all classifiers, Random Forest produced the best overall results across the tested feature set sizes. J48 produced competitive results specifically for feature sizes of 28 and 35 but underperformed on smaller sets.

Conclusion

This paper presented a hybrid feature selection mechanism using Naive Bayes and an Improved BAT Algorithm to enhance Intrusion Detection Systems. By reducing the feature space of the CICIDS2017 dataset, we demonstrated that machine learning classifiers could achieve high accuracy with optimized feature subsets. Future work will concentrate on multi-classification improvements and further hybridization of neuro-genetic models to address the limitations in detecting Web and Infiltration attacks.

References

[1] J. Wang and G. Beni, "Distributed computing problems in cellular robotic systems," in Proc. IEEE/RSJ Int. Workshop Intell. Robots Syst. (IROS), Tsukuba, Japan, 1989, pp. 819–826.

[2] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.

[3] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993.

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[5] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. 4th Int. Conf. Inf. Syst. Secur. Privacy (ICISSP), Portugal, Jan. 2018, pp. 108–116.

[6] K. Tamura, M. Sakiyama, and I. Arizono, "Ant system employing individual memories for the traveling salesman problem," IEEJ Trans. Electr. Electron. Eng., vol. 16, no. 1, pp. 118–125, 2021.

[7] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," in Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), vol. 284, J. R. González et al., Eds. Berlin, Germany: Springer, 2010, pp. 65–74.

[8] Canadian Institute for Cybersecurity, "CICIDS2017 Dataset," University of New Brunswick, 2018. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html.

[9] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," Erciyes University, Turkey, Tech. Rep. TR06, 2005.

[10] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[11] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[12] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.

[13] M. Dorigo, M. Birattari, and T. Stutzle, "Ant colony optimization," IEEE Computational Intelligence Magazine, vol. 1, no. 4, pp. 28–39, 2006.

[14] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, Mar. 2003.

[15] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," Machine Learning, vol. 29, no. 2, pp. 131–163, 1997.

[16] S. Mirjalili, S. M. Mirjalili, and X.-S. Yang, "Binary bat algorithm," Neural Computing and Applications, vol. 25, no. 3, pp. 663–681, 2014.

[17] D. E. Denning, "An intrusion-detection model," IEEE Transactions on Software Engineering, vol. SE-13, no. 2, pp. 222–232, 1987.

[18] J. Kennedy and R. Eberhart, "Particle swarm optimization," in Proc. ICNN'95 - Int. Conf. Neural Networks, vol. 4, Perth, WA, Australia, 1995, pp. 1942–1948.

[19] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016.

[20] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining. Boston, MA: Kluwer Academic Publishers, 1998.

[21] G. H. John and R. Kohavi, "Wrappers for feature subset selection," in Artificial Intelligence, vol. 97, no. 1-2, pp. 273-324, 1997.

Citation: *Naga Charan Nandigama. "A Hybrid Approach for Feature Selection Analysis on the Intrusion Detection System Using Naive Bayes and Improved BAT Algorithm". Research Journal of Nanoscience and Engineering. 5(1), 2021, pp 15-19.. DOI: https://doi.org/10.22259/2631-5591.0501003*