Global Journal of Engineering Innovations & Interdisciplinary Research

Original Article

Anomaly Detection in Network Traffic Using Machine Learning Techniques

T Sai Lalith Prasad[1], Beeram Aditya[2], Bodige Likhitha[2], G A Asta Govardhan Reddy[2]

[1]Assitant Professor, Department of Artificial Intelligence and Data Science, Vignan Institute of Technology and Science, Hyderabad, India

[2]UG Student, Department of AI&DS, Vignan Institute of Technology and Science, Hyderabad, India

Correspondence

T. Sai Lalith Prasad

Assistant Professor, Department of Artificial Intelligence and Data Science, Vignan Institute of Technology and Science, Hyderabad, India

- Received Date: 25 May 2025
- Accepted Date: 15 June 2025
- Publication Date: 27 June 2025

Keywords

Machine Learning, Network Anomaly Detection, CICIDS2017, Naive Bayes, QDA, Random Forest, ID3, AdaBoost, MLP, KNN

Abstract

It has been nothing less than exponential growth in the last two decades, and with this growing Internet has come unprecedented connectivity, significantly increasing the number of cyberattacks. Zero-day attacks have always challenged traditional signature-based detection techniques, which is why anomaly-based detection techniques have become increasingly important for identifying any anomalies in normal network behavior. Key features were selected using the Random Forest Regressor. Seven machine learning algorithms are tested in this experiment.

Introduction

The Internet revolution has transformed people's communication as in the two last decades, this development has never recorded such wide reach interconnecting across the whole globe for almost billion users of Internet. That trend in growing interconnectivity brought along significant security challenges from hackers characterized by frequency and complexities, which continue rising. Consequently, digital protection needs have made advanced threat detection their priority concerns individually and in terms of their entities.

Traditional intrusion detection systems primarily rely on signature-based approaches. These detection systems compare incoming data with a database of known attack patterns. While such systems are very efficient in identifying already encountered threats, they fail miserably when it comes to zero-day attacks. Zero-day attacks involve newly discovered vulnerabilities for which there exists no corresponding signature yet. Signature-based systems are, thus, useless when dealing with such threats.

Emergence of anomaly-based detection methods is only an approach toward new systems that would overcome the disadvantages of traditional signature-based systems. The new systems can identify possible threats based on how systems differ from their typical behavior rather than attack signatures. It is quite useful in dynamic environments where new emergent threats occur at every turn. Thus, the field is a very significant area of research in modern cybersecurity research.

This area of anomaly-based detection systems via machine learning could be applied concerning the major advancements in combating the cybersecurity threat. Machine learning is trained on the network traffic that looks for patterns which can, then, help make a difference in normal and harmful activities. By this respect, learning novel threats means being able to amplify the effectiveness with which the virtual environment is secured.

A dataset for the collection of network traffic is provided, known as CICIDS2017. The comparison on different machine learning algorithms used for the detection of network anomalies as well as the evaluation based on these algorithms will be demonstrated.

Seven algorithms had been tested of which were analyzed namely: Naive Bayes, Quadratic Discriminant Analysis, Random Forest, K-Nearest Neighbors, many others, determining the performance on the accurate type of attacks for identification. As of now results such as rankings that illustrate how the performance works have been produced, and therefore such results mark how machine learning is much in need for these security measures that are put across against all those emerging threats.

Related works

Kostas et al. [1] developed a big proposal on the anomaly detection problem in networks through various approaches by using machine learning methodology. In the paper, the authors exposed the inefficiency of the classic signature-based system in the fight against zero-day attacks and advocated the anomaly-based approach for detection purposes. It

Citation: Prasad TSL, Beeram A, Bodige L, Reddy GAAG. Anomaly Detection in Network Traffic Using Machine Learning Techniques. GJEIIR. 2025;5(4):087.

evaluates the accuracy of various detection types by executing several machine learning algorithms, namely Random Forest and K-Nearest Neighbors, in addition to the results of its implementation using the CICIDS2017 dataset.

Leung and Leckie et al. [2] present an unsupervised approach toward network intrusion detection by using the clustering methods. Their technique enables them to find the anomalous behaviors of the networks without the need to use predefined labeled datasets and challenges like dynamic attack patterns. Their paper at the Australasian Conference on Computer Science showed the potential effectiveness in detecting new attacks while diminishing the reliance on predefined signatures that would make their work more useful for real-world applications.

Sharafaldin et al. [3] developed the CICIDS2017 dataset. It soon proved to be a benchmark for many state-of-the-art intrusion detection systems during the performance testing and comprised mixed traffic in which a variety of types of attacks are blended in with the usual flow of network activity to make an evaluation tool sturdier as well as much more relevant toward research toward development of modern anomaly detectors to cope with a plethoric diversity of complex threats.

Thomas et al. [4] critically analyzed the utility of the DARPA dataset in evaluating intrusion detection systems. Older types of attacks and synthetic network traffic cannot reflect the reality of situations; thus, this study is pushing for newer datasets and metrics to represent today's challenges in cybersecurity and is recommending a refresh of standard practice for performance evaluation.

Ahmed et al. [5] had conducted a survey on methods for network anomaly detection, discussing a deep review of existing techniques and their application to the latest challenges in cybersecurity. The detection approach classification is divided by the survey into signature-based, anomaly-based, and hybrid models, and the advantages/disadvantages of these categories will be discussed. It means that machine learning is important for enhancing network security as it can halt attacks from sophisticated mechanisms, hence marking the importance of such technologies towards emerging threats.

Over these studies, this paper points toward the advancement in anomaly detection systems with a hint of weakness behind traditional methods as well as appreciating the benefit of using a highly dimensional source of data similar to CICIDS2017. Research currently is starting to indicate that combination of latest methodologies and high quality data is, in fact the requirement to build intrusion detection system more effective, in today's dynamically changing world of cybersecurity.

Proposed Methodology

Data preprocessing and preparation

This has highly improved the strength of this research based on the newest version of the CICIDS2017 dataset. Cleaning at the preprocessing stage removes missing values and inconsistencies, uniform formats, and extraction of key features such as packet size, flow durations, and protocol types from raw network traffic data. The processed data is divided into training and testing subsets to train and validate the model for proper assessment of the performance of the model against unseen data.

Feature Selection Using Random Forest Regressor

The random forest regressor performs feature selection. It ranks all the features that are available in the dataset in order to infer the vital ones for proper differentiation between normal traffic and malicious traffic. Among such selected important features are total forward packet length and standard deviation of packet length plus count of some flags like SYN or ACK flags.

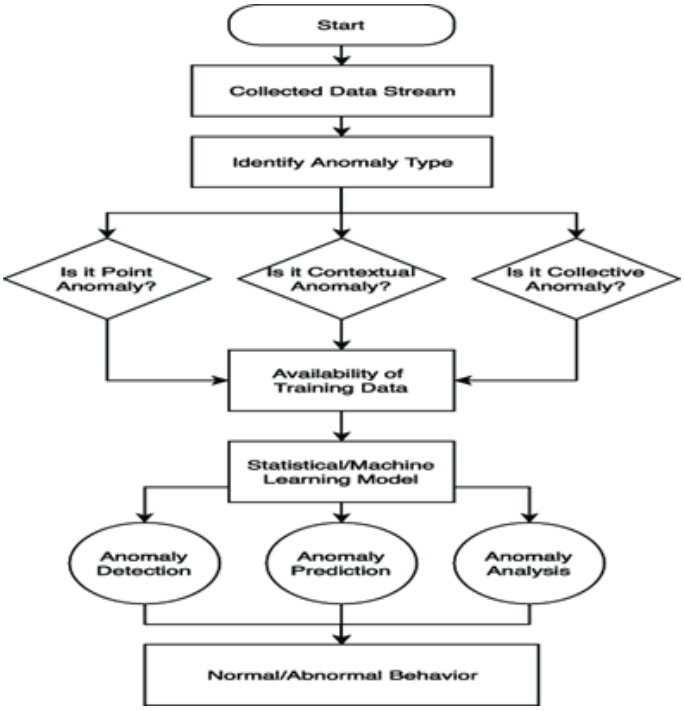
Application of Machine Learning Algorithms

It uses seven algorithms of machine learning for detecting anomalies, namely Random Forest, K-Nearest Neighbors (KNN), Decision Tree (ID3), Naive Bayes, Quadratic Discriminant Analysis (QDA), Multi-Layer Perceptron (MLP), and AdaBoost. The above algorithms operate on labeled data; they judge the accuracy of these algorithms through precision, recall, and F1-score metrics.

Model Analysis and Evaluation

It would be the comparative accuracy, precision, recall, and F1-score in the model for the analysis and the evaluation that largely guides the most suited algorithm that would be needed to deploy. A review of the confusion matrix that summarizes the true positives, false positives, true negatives, and false negatives will help determine the algorithm that results in the best possible detection rates balanced by the minimum possible false alarms.

Through the use of data-driven insight into cutting-edge machine learning, the book presents this systematic approach towards comprehensive and effective means of Network Anomaly Identification.



DATA SET

The CICIDS2017 dataset was used in this experiment. The CICIDS2017 dataset is developed by the Canadian Institute for Cybersecurity and is an updated, full version of modern network traffic comprising normal and malicious activities. To simulate the realistic network, devices were running various versions of Windows, Mac, and Linux. It includes more than five days of different network traffic with benign activity and 15 types of varied attacks, making it highly valuable for the training and testing of models to detect anomalies in the network through machine learning.

The distinguishing feature of this CICIDS2017 dataset is that it allows supervised machine learning. Data instances are classified as either benign or one of several categories of attacks like Distributed Denial of Service, Port Scans, web-based attacks in the form of SQL Injection and Cross-Site Scripting, among others. From raw network data, a high-dimensional feature set was extracted. Some of the features extracted were packet length, flow duration, and flag statistics. These are the significant features in the process of feature selection and are very informative for anomaly detection. Moreover, the dataset labels traffic as encrypted and non-encrypted, hence further enhancing its utility.

There are a lot of advantages related to the CICIDS2017 dataset; however, there are many disadvantages as well. For instance, it is humungous in its nature and therefore creates storage and processing problems. Raw network traffic files have a size of around 48 GB in the PCAP format and more than 1 GB of CSV files results after processing. This can be relaxed by using various data preprocessing techniques such as cleaning, normalization, and feature selection for ease of handling and preparation for analysis. This dataset will be characterised by high fidelity and diversity, and it will be an advanced resource to support the development and testing of the models that would eventually be used for the detection of anomalies in the networks.

Machine Learning Algorithms

Several strengths about anomaly detection in network traffic have been seen with the algorithms developed within this study. In particular, Random Forest is one of the ensemble learning approaches that aggregates multiple decision trees to obtain better classification accuracy and prevent overfitting through the building of each tree from a random subset of features. It can filter out select inputs by computing in training on the fly, therefore pointing to most distinguishing features, which most probably classify the traffic into being either normal or malicious.

Other basic algorithms used in the paper are the KNN algorithms. KNN is known by its simplicity and for being robust: it is classifier for instances as a function of the majority class in the feature space among the neighbors. Such algorithms are very efficient in processing high-dimensional data, such as network traffic logs, where one can use the size of a packet and the duration of a flow to indicate possible anomalies. Again, the KNN algorithm is highly dependent on both the chosen "K" and the distance metric used, so it requires careful tuning for optimum performance.

Other algorithms used include Naive Bayes, as it is simple and efficient, and AdaBoost; the iterative algorithm emphasizes misclassified instances improving weak classifiers. MLP is an artificial neural network with depth and support of non-linearity to detect more complex anomalies. QDA and ID3 have also been employed since they may model nonlinear relationships and make boundaries interpretable in decision.

This work utilizes them for the purpose of enabling the comparison of strengths of various diverse algorithms capable of detecting different sorts of network attacks. Therefore, all the analysis as above gives all the details in terms of a comparison of relative strength and suitability among algorithms for other anomaly detection contexts.

Types of Attacks

The CICIDS2017 dataset provides several types of network attacks to try to replicate the realistic cyber threats that could be constructed to create a comprehensive base for training machine learning models in the anomaly detection domain. Hence, the diversity of attacks created from various techniques attacking different parts of the network infrastructure is contributing to making the dataset more applicable for conducting robust cybersecurity research.

Denial of Service (DoS) Attacks: Denial-of-Service DoS attacks are based on the objective of overwhelming a target system or network, and they are guaranteed to ensure that access is impossible for legitimate users. Samples include: HULK, that forces servers to keep getting HTTP requests; GoldenEye, a multi-threaded, python-based attack oriented toward resource depletion, and others such as Slowloris, that rely on connection management by keeping idle connections open over time and with the purposes of continuing resource exhaustion.

Distributed Denial of Service (DDoS) Attacks: DDoS differs from DoS because there are multiple machines attacking a target at the same time. DDoS attacks are aimed at consuming a victim's bandwidth or processing power. LOIC is a traffic generator that can send high volumes of traffic from various sources, simulating the scenario of large attacks intended to overwhelm network infrastructure.

PortScan Attacks: In PortScan attacks, the network is scanned in order to detect available open ports so that the ones which can be accessed may be exploited. Scanning tools such as Nmap carry out systematic scans; therefore, the possibilities of finding any vulnerabilities in the target system cannot be ruled out. PortScans are largely used as reconnaissance, and hence the attackers may obtain information related to the structure of the network and the services available.

Brute Force Attacks: Generally, Brute Force attacks are systematic attempts to try many different usernames and passwords with the aim of cracking login credentials. Such events are illustrated within the data set. FTP-Patator illustrates attempts at unauthorized access to FTP servers, while SSH-Patator illustrates attempts at unauthorized access into SSH-enabled devices. Of such nature, the attacks result in numerous login attempts over a span of time, which usually linger in network logs.

Web-Based Attacks: Web-Based exploits web application flaws. SQL Injection makes changes to queries that access and modify data by the database; Cross-Site Scripting makes use of bad scripts injected in web pages executed on users' browsers. Brute force against web application login attempts to breach unauthorized access in web interfaces

Botnet Attacks: This attack is founded on a compromised device or a network of botnets, with which an attacker can use in order to facilitate malicious activities, including spamming and DDoS attacks. The dataset features botnet-traffic simulation representing coordinated malicious acts by infected devices.

Infiltration Attacks: The infiltrators gain access to the compromised systems with malware through an email or downloaded files within the network. They collect intelligence, find internal systems, and then exploit the weaknesses once inside. All these examples are data sets that show that after infiltrations, more unauthorized activities follow.

Heartbleed: Heartbleed attacks are a vulnerability in the OpenSSL library that allows an attacker to read data directly from a server's memory. Such an attack has been included in this dataset in which crafted requests find sensitive information, such as a password or encryption key.

Experimental Result

Graphical Representation for features Importance

This graphical representation is very important in the result. Models of anomaly detection appear in a graphic and intuitive shape in terms of evaluating its performance. This is the most effective tool for comparison to see which algorithm in the machine learning is more precise, KNN, Random Forest, or Naive Bayes. This kind of visual comparison is very transparent to know how the models could be relatively successful in detecting anomalies in the given data set.

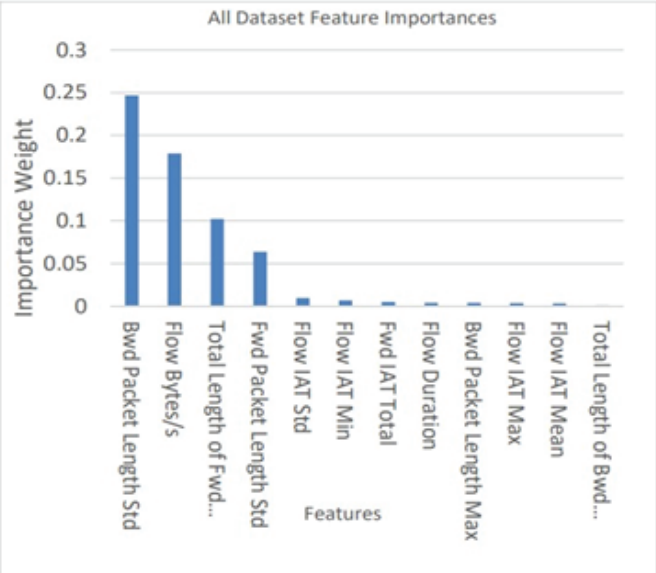
A detailed view of the prediction outcomes of the model can be viewed in the form of a heatmap known as the confusion matrix. It exhibits the distribution of true positives, true negatives, false positives, and false negatives, thereby showing valuable insights regarding the strengths and weaknesses of the model. In this regard, visualization is quite useful in the identification of areas where the model is good or bad, for example, missed detections or false alarms.

The other important graphical tool is precision-recall curves. Precision-recall curves are used to find out the trade-off for each model concerning the precision and the recall. In such cases where an individual has a problem of imbalanced data sets, they greatly require spotting the anomaly in its class correctly. Precisions vs. Recalls might show much more insights about what might occur as a model approaches various conditions.

This is another feature importance, which might be visualized as bar plots-it will display that which attributes most influence the predictions of the model. This kind of visualization might be helpful in understanding that which features are most influential in the differentiation between normal and anomalous network traffic.

Feature Selection

This is one of the basic steps in developing machine learning models for anomaly detection that helps simplify the complexity of the dataset and thus improves the precision and efficiency of the model. The CICIDS2017 dataset is of very high dimensions; therefore, feature selection was conducted to filter the most essential features for appropriate anomaly detection based on the relevance by the Random Forest Regressor, considering importance and ranking features in classification problems.



The most significant features that were more prominent include the total number of packets, flow duration, mean packet size, and standard deviation of the size of the packets moving forward. This reduction to the top-ranked features within the dataset meant that redundant or less impactful features were eradicated; consequently, it reduced the complexity of the dataset but improved the performance and efficiency of machine learning models.

Meaningful feature selection was useful for optimizing the applied model that could be used for conducting the study. Most retained meaningful attributes could help better performing machine learning algorithms for accurate anomaly detection. Advantages in regard to insights acquired from Random Forest Regressor regarding the conducted study are valuable to make sure the most influential features are used, thereby boosting the overall performance of anomaly detection models.

Model Accuracy

The model reflects the results of accuracy through an evaluation of the CICIDS2017 dataset, and that is where one can clearly find a relative effectiveness in the detection in anomalies with different algorithms applied by machine learning. Among all these proposed algorithms, the ranker for this experiment was KNN at a rank of 97%. After it ranked the rest with Random Forest and AdaBoost by scoring 94% in the accuracy of detecting anomalies. Also, the ID3 algorithm for the Decision Tree was going pretty well with 95% accuracy. Naive Bayes and QDA comparatively performed averagely with the same accuracy of 86%. MLP was at a very low level with only accuracy of 83%. Perhaps this could be because of the fact that MLP is highly sensitive to the high dimensional feature space within its dataset.

References

1. K. Kostas, Anomaly Detection in Networks Using Machine Learning. Research Proposal, March 23, 2018.
2. Miniwatts Marketing Group, "Internet Growth Statistics," 2018. Available online: https://www.internetworldstats.com/emarketing.htm [Accessed August 26, 2018].
3. K. Leung and C. Leckie, "Cluster-based unsupervised anomaly detection for network intrusion detection," presented at the Australasian Conference on Computer Science, 2005, pp. 333-342. Published by the Australian Computer Society, Inc.
4. Sharafaldin, A. Gharib, A. H. Lashkari, and A. A. Ghorbani, "Development of a reliable benchmark dataset for intrusion detection," Software Networking, vol. 1, no. 1, pp. 177-200, 2017.
5. Massachusetts Institute of Technology, Lincoln Laboratory, "DARPA Intrusion Detection Evaluation Dataset 1998." Available at: https://www.ll.mit.edu/rd/datasets/1998-darpa-intrusion-detection-evaluation-data-set. [Accessed August 5, 2018].
6. C. Thomas, V. Sharma, and N. Balakrishnan, "Evaluating the utility of the DARPA dataset for testing intrusion detection systems," presented at the conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security, vol. 6973, 2008.
7. Gharib, I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Framework for evaluating intrusion detection datasets," presented at the International Conference on Information Science and Security (ICISS), 2016, pp. 1-6. Published by

IEEE.

8. University of California, Irvine, "KDD Cup 1999 Dataset." Available online: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. [Accessed August 5, 2018].
9. Özgür and H. Erdem, "Reviewing the usage of the KDD99 dataset for intrusion detection research between 2010 and 2015," PeerJ Preprints, vol. 4, article e1954v1, 2016.
10. Center for Applied Internet Data Analysis (CAIDA), "OC48 Peering Point Traces Dataset." Available online: https://www.caida.org/data/passive/passive_oc48_dataset.xml. [Accessed August 6, 2018].
11. M. Ahmed, A. N. Mahmood, and J. Hu, "Survey of techniques for detecting network anomalies," Journal of Network and Computer Applications, vol. 60, pp. 19-31, 2016.
12. University of New Brunswick, Canadian Institute for Cybersecurity, "NSL-KDD Dataset." Available online: http://www.unb.ca/cic/datasets/nsl.html. [Accessed August 6, 2018]. R. Bhallamudi et al., "Deep Learning Model for Resolution Enhancement of Biomedical Images for Biometrics," in Generative Artificial Intelligence for Biomedical and Smart Health Informatics, Wiley Online Library, pp. 321–341, 2025.
13. R. Bhallamudi et al., "Artificial Intelligence Probabilities Scheme for Disease Prevention Data Set Construction in Intelligent Smart Healthcare Scenario," SLAS Technology, vol. 29, pp. 2472–6303, 2024, Elsevier.
14. R. Bhallamudi, "Improved Selection Method for Evolutionary Artificial Neural Network Design," Pakistan Heart Journal, vol. 56, pp. 985–992, 2023.
15. R. Bhallamudi et al., "Time and Statistical Complexity of Proposed Evolutionary Algorithm in Artificial Neural Networks," Pakistan Heart Journal, vol. 56, pp. 1014–1019, 2023.
16. R. Krishna et al., "Smart Governance in Public Agencies Using Big Data," The International Journal of Analytical and Experimental Modal Analysis (IJAEMA), vol. 7, pp. 1082–1095, 2020.
17. N. M. Krishna, "Object Detection and Tracking Using YOLO," in 3rd International Conference on Inventive Research in Computing Applications (ICIRCA-2021), IEEE, Sept. 2021, ISBN: 978-0-7381-4627-0.
18. N. M. Krishna, "Deep Learning Convolutional Neural Network (CNN) with Gaussian Mixture Model for Predicting Pancreatic Cancer," Springer US, vol. 1380-7501, pp. 1–15, Feb. 2019.
19. N. M. Krishna, "Emotion Recognition Using Skew Gaussian Mixture Model for Brain–Computer Interaction," in SCDA-2018, Textbook Chapter, ISBN: 978-981-13-0514, pp. 297–305, Springer, 2018.
20. N. M. Krishna, "A Novel Approach for Effective Emotion Recognition Using Double Truncated Gaussian Mixture Model and EEG," I.J. Intelligent Systems and Applications, vol. 6, pp. 33–42, 2017.
21. N. M. Krishna, "Object Detection and Tracking Using YOLO," in 3rd International Conference on Inventive Research in Computing Applications (ICIRCA-2021), IEEE, Sept. 2021, ISBN: 978-0-7381-4627-0.
22. T. S. L. Prasad, K. B. Manikandan, and J. Vinoj, "Shielding NLP Systems: An In-depth Survey on Advanced AI Techniques for Adversarial Attack Detection in Cyber Security," in 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS), IEEE, 2024.
23. S. Sowjanya et al., "Bioacoustics Signal Authentication for E-Medical Records Using Blockchain," in 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), vol. 1, IEEE, 2024.
24. N. V. N. Sowjanya, G. Swetha, and T. S. L. Prasad, "AI Based Improved Vehicle Detection and Classification in Patterns Using Deep Learning," in Disruptive Technologies in Computing and Communication Systems: Proceedings of the 1st International Conference on Disruptive Technologies in Computing and Communication Systems, CRC Press, 2024.
25. V. P. Krishna and T. S. L. Prasad, "Weapon Detection Using Deep Learning," Journal of Optoelectronics Laser, vol. 41, no. 7, pp. 557–567, 2022.
26. T. S. L. Prasad et al., "Deep Learning Based Crowd Counting Using Image and Video," 2024.