OPEN ACCESS

ISSN: 2996-4954

Rajiv C, et al. (2022)100014
DOI: 10.71010/2996-4954/JCETAI/102

Research Article

Journal of Contemporary Education Theory & Artificial Intelligence

Leveraging Big Datasets for Machine Learning-Based Anomaly Detection in Cybersecurity Network Traffic

Rajiv Chalasani[1*], Mukund Sai Vikram Tyagadurgam[2], Venkataswamy Naidu Gangineni[3], Sriram Pabbineedi[4], Mitra Penmetsa[5], Jayakeshav Reddy Bhumireddy[6]

[1]Sacred Heart University, Rajivchalasani555@gmail.com
[2]University of Illinois at Springfield, t.vikkhram@gmail.com
[3]University of Madras, Chennai, vgangineni@gmail.com
[4]University of Central Missouri, sreeram7766@gmail.com
[5]University of Illinois at Springfield, mitravarma.penmetsa@gmail.com
[6]University of Houston, jayakeshav10807@gmail.com

[*]**Correspondence:** Rajiv Chalasani

Citation: Rajiv C, Mukund Sai VT, Venkataswamy Naidu G, Sriram P, Mitra P, et al. (2022) Leveraging Big Datasets for Machine Learning-Based Anomaly Detection in Cybersecurity Network Traffic. J Contemp Edu Theo Artific Intel: JCETAI/102.

Received Date: 05 November, 2022; **Accepted Date:** 16 November, 2022; **Published Date:** 23 November, 2022

Abstract

In order to improve cybersecurity and stop criminal activity, network traffic anomaly detection is essential. Anomaly detection is an essential part of cybersecurity, which is necessary to find complex and until undiscovered network threats that frequently evade detection techniques based on signatures and heuristics. This study proposes a comprehensive machine learning framework employing the Random Forest (RF) algorithm, combined with an advanced data preprocessing pipeline encompassing data cleaning, encoding of categorical features, class balancing using SMOTE, feature selection, and normalization to enhance model input quality. The method is tested on the well-known CICIDS2017 dataset, which records a wide variety of current cyberattacks and safe network activity. The suggested RF model performs exceptionally well, attaining 99.88% accuracy, precision, recall, and F1-score. According to comparative findings, the RF model performs much better than baseline methods like K-Nearest Neighbors and Linear Regression, which obtained far lower evaluation metrics. In increasingly complex digital settings, these findings highlight the model's scalability, durability, and applicability for real-time intrusion detection, which helps to create cybersecurity defenses that are more resilient and adaptable.

Keywords: *Anomaly Detection, Network Intrusion Detection, Cybersecurity, CICIDS2017 Dataset, Network Traffic Analysis, Cybersecurity.*

I. Introduction

Cybersecurity involves a wide range of issues, such as malware analysis, attribution, incident response, and intrusion detection. The identification of harmful network activity, in its broadest sense, is our main focus here. In the past, malware signatures and heuristic-based rules have been the mainstays of intrusion detection software, antivirus software, and related products. Matching approaches such as this work well when security software makers are aware of such assaults beforehand [1]. A separate strategy is needed to identify harmful events on a network that have not yet been identified or for which detection criteria are not yet in place. Leveraging big dataset collected from large-scale network traffic enables the development of more robust and scalable anomaly detection techniques [2]. Here, it demonstrates the modelling of a network's "normal" behaviour using unsupervised ML to find abnormalities on the network and scoring deviations from this baseline. Anomaly detection is essential for spotting hidden dangers that go unnoticed by conventional defences. By focusing on behavioural deviations rather than predefined signatures, this approach enhances the adaptability and resilience of cybersecurity systems.

Detecting abnormalities in data is crucial and this process is called anomaly detection. This field is exciting because it helps us discover rare and interesting patterns in data [3]. This method, which goes by several names in statistics, ML, exception mining, novelty identification, and outlier detection, has drawn a lot of interest from the scholarly community [4]. Anomalies are valued since they highlight important but infrequent events and may lead to urgent actions in different types of industries, in the field of network security, detecting anomalies is necessary to discover unusual traffic that could indicate cyber-attacks or data breaches.

This problem involves detecting when the rules of network traffic are no longer normal [5]. Usually, the traffic of network attacks does not resemble normal internet traffic. Using this principle, the traffic detection algorithm can detect any unwanted traffic [6]. At the moment, the main techniques used for detecting network traffic anomalies are: statistical analysis, ML and neural network are some of the methods used in artificial intelligence.

The use of big data analytics makes ML-based anomaly detection systems perform better and scale well. These solutions can analyse large amounts of real-time data from different networks to build a broad view of how things work in different environments. By leveraging high-dimensional feature spaces [7], Using data from the network can help models recognize all the details of interactions, increasing the accuracy of detection and reducing false positives. Additionally, you can use techniques to choose the most important features which makes analysis simpler and clearer.

A. Motivation and Contributions of the Study

The reason for this study is the increasing number and complexity of cyber-attacks focusing on cloud-based networks and corporate systems. It is difficult for traditional intrusion detection systems to handle new attack methods and unbalanced data, causing them to give many false alerts and have limited capacity to grow. With the increasing reliance on real-time data transmission and the expanding surface for attacks such as DDoS, botnets, and web-based intrusions, there is a critical need for intelligent, adaptive, and efficient anomaly detection mechanisms. In order to overcome these obstacles, this study will use ML, more especially the RF algorithm, in conjunction with sophisticated data preparation methods to create a reliable IDS. The goal is to enhance detection accuracy, minimize misclassifications, and ultimately contribute to more secure and resilient network environments. This study makes the following key contributions:

- Developed a robust data pre-processing pipeline, including cleaning, encoding, SMOTE for class balancing, feature selection, and normalization, to enhance model input quality for anomaly detection.
- Efficient Use of Random Forest Model: Implemented and optimized the RF algorithm for intrusion detection, demonstrating its capability to accurately classify both benign and malicious traffic in high-dimensional, imbalanced network data.
- Performance Validation on Realistic Dataset: Evaluated the model using the dataset from CICIDS2017, which covers a variety of contemporary cyberattacks, to ensure realistic and reliable performance assessment using important indicators including F1-score, recall, accuracy, and precision.
- Comparative Analysis with Baseline Models: Comparative research was carried out to demonstrate the superior performance and usefulness of the suggested technique in cybersecurity situations versus conventional ML models like LR and KNN.

B. Justification and Novelty of paper

The novelty of the proposed approach lies in its integration of a robust ML model RF with a well-structured data pre-processing pipeline, optimized specifically for handling the complexities of modern network traffic. By leveraging the comprehensive CICIDS2017 dataset and addressing common challenges such as data imbalance through SMOTE, as well as applying rigorous feature selection and normalization techniques, the methodology ensures high-quality input for model training. Unlike traditional techniques, the RF model's ensemble learning mechanism provides enhanced generalization, enabling accurate detection of both common and sophisticated cyber threats. This approach is further justified by its consistent and near-perfect performance across key evaluation metrics, as evidenced in the experimental results. The combination of advanced pre-processing, strategic model selection, and empirical validation supports the solution's efficacy and feasibility in detecting cybersecurity anomalies in the actual world.

C. Organization of the paper

The structure of the paper is as follows Section II examines relevant research on cybersecurity anomaly detection. The machine learning models and technique are explained in Section III. Model comparisons and experimental findings are shown in Section IV. The study is concluded in Section V, which also addresses potential avenues for further research.

II. Literature of Review

In this section, they will highlight ML methods that are used to spot unusual activity in network traffic, including techniques such as supervised learning, DL and those that combine them. By analysing current and large datasets, the studies have shown accurate results in identifying and sorting network anomalies.

Lin et al. (2019) The security of the network is ensured by designing and implementing a network anomaly detection system that makes use of DL techniques. A DNN is constructed using an LSTM model, and its performance is subsequently improved by adding an AM. Using the SMOTE technique and an improved loss function, the team was able to rectify the dataset's classified imbalance. Their model's accuracy in classification is 96.2% which is more accurate when compared to other machine learning models [8].

Atefi et al. (2019) This study aims to analyze data anomalies for the intrusion detection system's classification function utilizing the most recent CICIDS-2017 dataset, which facilitates intrusion detection evaluation. The data in this study was subjected to anomaly analysis and classification utilizing DNN from DL and KNN for ML. To evaluate how well the ML and DL techniques work in the classification portion of the findings, the MCC is utilized. The correctness classifier has a score range, and DNN's accuracy of 0.9293% was somewhat higher than KNN's of 0.8824% [9].

Alrashdi et al. (2019) The Anomaly Detection-IoT system, which employs a RF algorithm to identify anomalous occurrences, is their recommendation for cybersecurity in smart cities. The suggested method may efficiently identify hacked IoT devices at dispersed fog nodes. It's tested their model using a contemporary dataset to demonstrate its correctness. their research shows that the AD-IoT can maintain a low FRP while reaching a maximum classification accuracy of 99.34%.[10].

Srivastava et al. (2019) In the past, intrusions in network traffic data were found using supervised learning techniques. However, not only has traffic expanded dramatically in recent years, but network threats are also evolving. Enhancements in detection methods are necessary to identify these novel forms of assaults. Researchers have thoroughly studied ML methods for identifying irregularities in network data. The public repositories now contain new datasets. then employed cutting-edge ML techniques based on feature reduction to find unusual patterns in the freshly supplied dataset. An impressive 86.15 percent accuracy rate has been attained [11].

Pattawaro and Polprasert, (2018) develop a network intrusion detection system that uses the XGBoost classification model, K-Means clustering, and feature selection to find anomalies. For the NSL-KDD dataset, they evaluate their proposed approach using the KDDTest+ dataset. Based on Detest+ data, their proposed two-cluster model achieves 84.41% accuracy, 86.36% detection rate, and 18.20% false alarm rate. Additionally, due to feature selection, their proposed model only trains to this performance level using 75 out of 122 features (61.47%), which is equivalent to models that use every feature [12].

Salman et al. (2017) investigates the process of both detecting and categorizing anomalies, rather than focusing just on the detection phase, as is typical in most recent studies. For the purpose of detecting and classifying various assaults, they have developed and evaluated learning models using a widely used

publicly accessible dataset. Two supervised ML methods, specifically RF and LR, have been employed. They demonstrate how comparable assaults can make classification less accurate even with flawless detection. Their findings show a detection accuracy of over 99% and a classification accuracy of 93.6%, while some assaults cannot be classified [13].

Table I provides a synopsis of relevant research on cybersecurity anomaly detection conducted in the last several years, highlighting methodologies, datasets, performance metrics, and identified limitations to guide future research directions.

Table I: Summary of the related work based on Anomaly Detection in Cybersecurity Network Traffic.

Reference	Methodology	Dataset	Performance	Limitations & Future Work
Lin et al., (2019)	LSTM with Attention Mechanism; SMOTE; Improved loss function	CSE-CIC-IDS2018	Accuracy: 96.2%	Needs more validation on real-time streaming traffic; model complexity could be improved.
Atefi et al., (2019)	KNN and DNN for anomaly classification	CICIDS-2017	MCC: DNN (0.9293), KNN (0.8824)	Limited to MCC as performance measure; further analysis on other metrics needed.
Alrashdi et al., (2019)	Random Forest-based AD-IoT for smart cities	IoT-focused modern dataset	Accuracy: 99.34%, Low FPR	Specific to smart cities; broader IoT deployment scenarios need exploration.
Srivastava et al., (2019)	Feature reduction + ML algorithms	Recently updated public dataset	Accuracy: 86.15%	Lower accuracy; needs improvement in detection of novel threats with scalable solutions.
Pattawaro and Polprasert, (2018)	Classification using XGBoost, K-Means clustering, and augmented reality feature selection	NSL-KDD (KDDTest+)	Acc: 84.41%, DR: 86.36%, FAR: 18.20%	Limited to 2 clusters; explore better clustering and real-time traffic use
Salman et al., (2017)	Linear Regression (LR), Random Forest (RF)	Public dataset (unspecified)	Detection: >99%, Categorization: 93.6%	Categorization challenges due to attack similarities; future work on multi-class classifiers.

III. Methodology

The proposed methodology follows a comprehensive, multi-phase approach for Figure 1 shows Network traffic irregularities were found using ML for cybersecurity. Gathering the CICIDS2017 dataset is the first step, which is pre-processed to ensure data quality through the removal of missing, duplicate, and redundant values, followed by data cleaning and encoding of categorical features. The SMOTE approach is used to correct class imbalance, and pertinent characteristics are chosen to increase model efficiency. Normalization is then performed to scale the feature values uniformly. A ML model, like RF, is trained to identify and categorized anomalies in the refined dataset after it has been partitioned into sections for testing and training. The model's performance is assessed using important performance indicators, including as precision, recall, and F1-score, accuracy which show how well the model detects intrusions in actual network settings. From using datasets to evaluating the final outcomes, the steps in the suggested method are depicted in Figure 1.

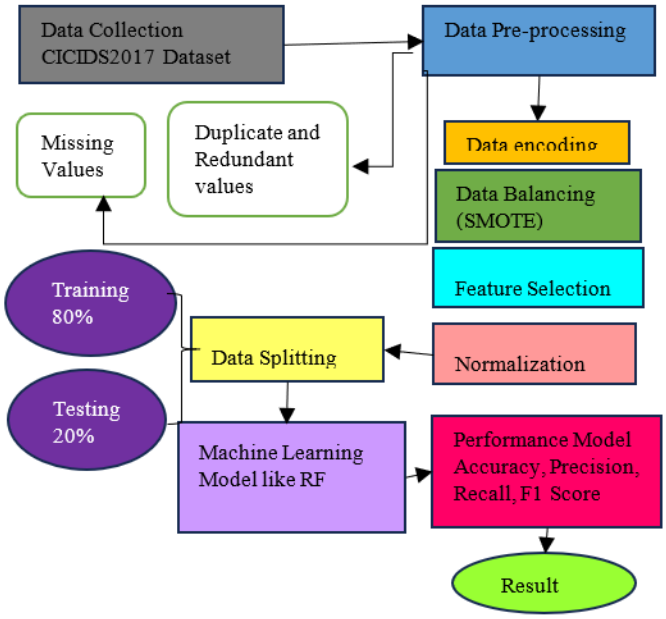

Figure 1: Data Flow Diagram for Anomaly detection.

A brief explanation of each data Figure 1 flow diagram phase is provided below:

A. Data Collection

A thorough and current benchmark dataset that is frequently used to assess intrusion detection systems is the CIC-IDS2017 dataset. It captures realistic network traffic generated using CICFlow Meter and includes labelled flow records based on key network attributes such as protocol types, attack classes, ports, date/time stamps, IP addresses (both sending and receiving), and so forth. The dataset covers a wide spectrum of contemporary cyberthreats, including as SQL injection, distributed denial of service, brute force, botnet assaults, and cross-site scripting, infiltration, and port scans. It is organized into eight CSV files and has 2,830,743 records in total, each characterized by 78 distinct features along with corresponding labels. Its richness and diversity make it well-suited for training and evaluating ML models in cybersecurity research.

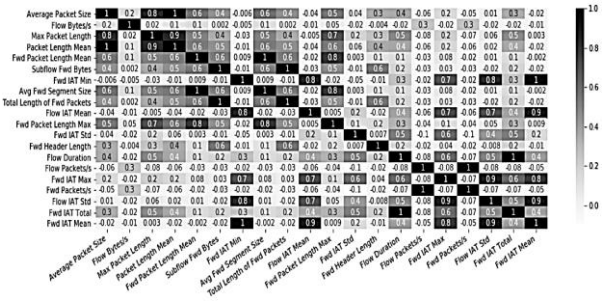

Figure 2: Correlation Matrix Heatmap on CICIDS2017 Dataset for Anomaly Detection.

Feature correlation analysis was performed on the CICIDS2017 dataset to identify interdependencies among network traffic attributes. The correlation matrix heatmap in Figure 2 helps choose the most important features and avoid repeating similar information. This stage boosts the model's performance by reducing multicollinearity and conserving just the relevant and associated features useful for spotting anomalies.

Figure 3 presents the dataset's class distribution, which demonstrates that the attack category has less network traffic than the benign category. The majority of the data contains benign traffic which represents 83% (235,289 records), while malicious network attacks are much less frequent, accounting for 17% (47,145 records). This unequal distribution makes it more difficult to create models that successfully spot minority-class network attacks in large datasets of mainly normal traffic.

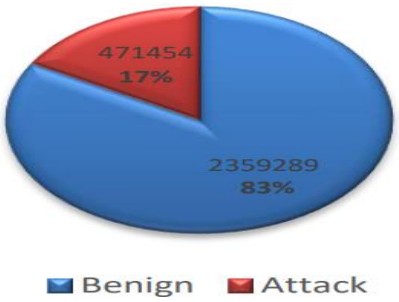

Figure 3: The Distribution in CICIDS2017 Dataset

B. Data Preprocessing

An important stage in data feeding into ML systems is preparation, particularly in fields like cybersecurity that rely on network traffic anomaly detection. Using the dataset from CICIDS2017. There are a few preprocessing steps, like cleaning the data, changing labels to a code system, dealing with imbalanced data using SMOTE, selecting features and normalizing them.

Missing Values: Find and throw out missing class labels and incomplete datapoints to maintain cleanliness and objectivity in the dataset.

Duplicate and Redundant Entries: Identify and eliminate duplicate or redundant records to reduce noise and improve dataset integrity.

C. Data encoding

Data encoding may be used to translate the labels into numerical values. Since our dataset is a binary classification issue, the set of possible values that the model may predict is also binary. In reality, the dataset labels are "0" and "1," with "0" meaning "Benign" and "1" meaning "Attack." Consequently, the model forecasts either '0' or '1' for each observation. In fact, encoding facilitates the model's comprehension of the input. The model would interpret 'Benign' as text instead of a label if the label were unencoded. The model can analyze the data more efficiently by using numbers for the labels when they are encoded [14].

D. Synthetic Minority Oversampling Technique (SMOTE) Technique

To create synthetic samples for the under-represented classes, the SMOTE method was utilized. [8], improving the dataset's balance.

SMOTE creates Interpolating between an instance of a minority class and one of its KNN yields fresh samples. The synthetic sample x_{new} is generated using the following formula in Equation (1):

$$x_{new} = x_i + \delta.(x_{nn} - x_i) \qquad (1)$$

Where:

- x_i is an original minority class instance,
- x_{nn} is one of its KNN,
- δ is a random value between 0 and 1,
- x_{new} is the recently produced synthetic sample.

This technique enhances the model's ability to recognize and classify underrepresented attack types by providing a more balanced training set.

E. Feature Selection

The term "Feature selection" describes the procedures that are employed to extract valuable data from a dataset by taking a subset of all available features and sorting them according to predetermined criteria [15]. In doing so, it helps reduce data processing size by removing redundant and excessive components.

F. Normalization with Min-Max

Data normalization techniques such as decimal scaling normalization, z-score, and min-max are utilized. Normalization processing a number of approaches, one of which is the Min-Max normalization technique, to numerical properties. By altering every attribute value within a specific range of [0, 1], it is crucial to improve the system's performance and efficacy. Equation (2), however, illustrates its anomalous affectability.

$$z = \left(\frac{x_i - \min(x)}{(\max(x) - \min(x))}\right) \qquad (2)$$

The data element is denoted by x_i, $\min(x)$ is the largest value in the data set, while $\min(x)$ is the lowest value, and Z is an arbitrary new value [16].

G. Data Splitting

The train and test set split were performed in a stratified manner with a split ratio of 80:20. A stratified split was used in order to preserve the training and testing sets' proportionate class distribution, which mirrored that of the original dataset.

H. Classification of Proposed Random Forest (RF) Model

RF is an ensemble learning technique that combines a large number of DT to generate predictions collectively [17]. Each DT in RF generates its own forecasts, which are then combined to provide the final prediction. Let's use X for the input features, Y for the target variable, and RF for the RF model. Equation (3) may be used to express the prediction of RF if there is N. DT present in the forest:

$$RF(X) = mode(Tree_1(X), Tree_2(X), \ldots., Tree_N(X)) \quad (3)$$

where $Tree_i(X)$ represents the i-th DT forecast. mode() yields the most common class label across all tree forecasts in a classification task. It is possible to substitute mode() in a regression job by averaging the predictions. A bootstrapped fraction of the training data is used to build each DT, and a random selection of attributes determines each node's predictions. The RF model can decrease overfitting and enhance generalization performance by combining predictions.

I. Performance Metrics

The evaluation is carried out using common classification measures, including confusion matrix, precision, recall, accuracy and F1-score. how well ML models use big datasets to find anomalies in cybersecurity network traffic. A popular technique for evaluating classification performance on test data with known real labels is a confusion matrix, summarizes the models' actual and expected classifications. Accurate interpretation of these metrics is crucial for effectively comparing model performance in anomaly detection tasks within cybersecurity contexts as discussed below:

- **True Positive (TP):** The result is deemed TP if the model detects an abnormality as such.
- **False Positive (FP):** The outcome is termed as FP if the model classifies a typical event as an abnormality.
- **True Negative (TN):** An anomaly is accepted as TN if the model determines that it is a typical instance.
- **False Negative (FN):** The outcome is identified as FN if the model detects a typical case as such.

1) Accuracy

The percentage of samples that are correctly categorized is known as accuracy. When classes are evenly distributed, accuracy is a useful metric [18]. Equation (4) may be used to determine the accuracy.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

2) Precision

The testing dataset's percentage of real anomaly samples out of all the anomaly samples that the detection model found is known as precision. Equation (5) allows for the calculation of precision.

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

3) Recall

The recall is the percentage of actual anomaly samples in the testing dataset compared to all anomaly samples. It is possible to calculate the recall using Equation (6).

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

4) F1-Score

The weighted average of recall and accuracy might be considered the F-measure. The Equation (7) displays the F-measure formula.

$$F - measure = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \quad (7)$$

5) ROC

The diagnostic capability of a classifier is demonstrated by the ROC curve, which displays TPR vs FPR in Equation (8,9).

$$TPR = \frac{TP}{(TP+FN)} \quad (8)$$

$$FPR = \frac{FP}{(FP+TN)} \quad (9)$$

These Performance Matrix are utilized for comparative analysis and evaluate the model performance for Anomaly Detection.

IV. Result Analysis and Discussion

In order to facilitate accurate evaluation and efficient processing, An Intel Core i9-13900K CPU (3.0 GHz), The high-performance PCs utilized for the research included Windows 11 Pro, 64 GB DDR5 RAM, and an NVIDIA RTX 4090 GPU with 24 GB VRAM. RF achieved exceptional results on the CICIDS2017 dataset, as seen in Table 3, with an precision, recall, accuracy and F1-score of 99.88%. This proves that the RF model can successfully detect any anomalies in a network. The model correctly discerned benign traffic from malicious attacks due to its use of ensemble learning which decreases the risk of overfitting and manages large data sets that have a big imbalance. These near-perfect scores suggest that these models can be applied to cybersecurity successfully and reliably detect anomalies.

Table II. Outcome of Random Forest (RF) model on the CICIDS2017 dataset.

Matrix	Random Forest (RF)
Accuracy	99.88
Precision	99.88
Recall	99.88
F1-Score	99.88

Table II presents the data using a bar graph, making the key figures of the RF model easier to understand. The bars consistently show strong performance on all the evaluated measures, underlining the ability of the model to identify and classify anomalous network activities. When we compare the Table with the bar chart, we clearly see how RF can improve cybersecurity by accurately detecting anomalies.

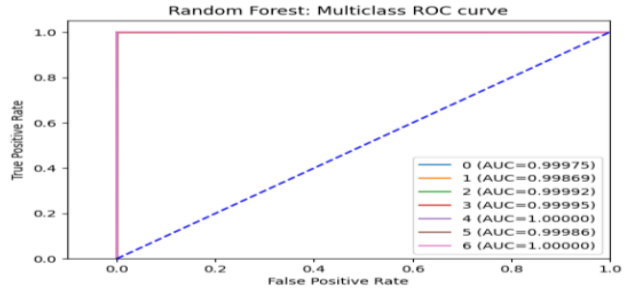

Figure 4: ROC Curve of Random Forest (RF).

Figure 4 highlights that the RF model performs strongly in multiclass settings by showing excellent performance in all classes. A low FPR and a high TPR are located around the upper-left corner, where the class-specific curves cluster. All of the AUC values fall between 0.99869 and 1.00000, meaning that all classes can be separated and classified almost perfectly. The model is the best option for accurate anomaly detection in cybersecurity as it successfully classifies different types of network data.

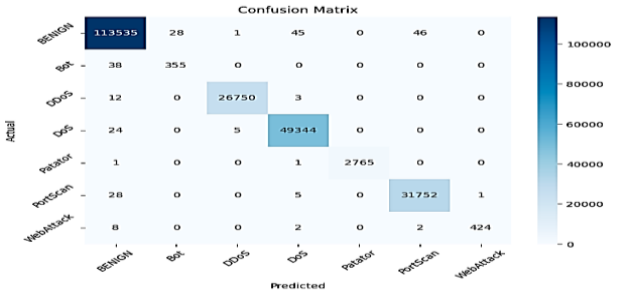

Figure 5: Matrix of Random Forest (RF).

Figure 5 displays The RF model's confusion matrix, which was trained on the CICIDS2017 dataset, correctly distinguishes between four categories of assaults—Bot, DDoS, DoS, and Web Attack—and benign attacks. The results show that a high number of true positives are found and only little misclassification happens, mainly for ordinary insects and severe attacks. The experiment was carried out with Scikit-learn and used Gini impurity as the criterion. The findings demonstrate that the model is accurate, precise and efficient in detecting cybersecurity anomalies.

A. Comparative Analysis

In this section, the focus is on highlighting the differences between anomaly detection techniques used in cybersecurity networks. Among machine and DL models, the RF model was found to be the best in F1-score, recall, accuracy, and precision in identifying irregularities in network traffic data. A comparison of LR is presented in Table III [19], KNN [20] and RF models are all utilized for cybersecurity-related network traffic anomaly detection.

The LR model's precision, recall, accuracy and F1-score are 64.11%, 71%, 64%, and 63%, in that order. However, the KNN model did much better, scoring 97% on each of the evaluation metrics. The RF model scored almost perfectly on 99.88% in terms of precision, recall, accuracy and F1-score. The following Table (Table III) summarizes how these methods measure up based on factors such as accuracy.

Table III: Comparative analysis for Anomaly Detection in Cybersecurity Network Traffic.

Model	Accuracy	Precision	Recall	F1-Score
Linear Regression (LR)	64.11	71	64	63
K Nearest Neighbours (KNN)	97	97	97	97
Random Forest (RF)	99.88	99.88	99.88	99.88

This RF model provides a number of important benefits for identifying anomalies in security network traffic. Such an architectural approach increases both the accuracy and reliability by using multiple decision trees to capture the many different and uncommon features found in vast datasets. Because it works well with high-dimensional data and can tolerate overfitting, the model can be trusted for real-time security against intrusions and DDoS attacks in dynamic cloud environments.

V. Conclusion and Future Scope

It is very important in modern times to detect changes in a network for security reasons. This study demonstrates how RF and a thorough preprocessing procedure enable efficient network anomaly detection using ML. Using the CICIDS2017 dataset, the proposed model demonstrated unusually high levels of precision, recall, accuracy and F1-score of 99.88%. With its capability to operate with large and unbalanced data, RF helps deliver swift and precise detection for many online threats. The findings support the idea that the model is useful in various cybersecurity contexts and can strengthen network security by preventing malicious attacks. It uses data that has been labeled and may not function properly when confronted with unknown attacks. Using neural networks can be costly in terms of power, so they are less used in high-speed networks.

Future studies could expand the proposed model by incorporating RNNs or GNNs, to better understand how traffic flows differ over time and between various points in the network. Since supervised and unsupervised learning have separate advantages, mixing them might enhance the detection of unknown and stealthy attacks. Adding real-time systems that are flexible and adding explainable approaches to the system will help the anomaly detection system be practical and trustworthy. Evaluating the model with more types of data and real systems confirms its ability to be applied widely.

References

1. B. J. Radford, B. D. Richardson, and S. E. Davis, "Sequence Aggregation Rules for Anomaly Detection in Computer Network Traffic," 2018.
2. V. Kolluri, "Vulnerabilities: Exploring Risks in AI Models and Algorithms," *IJRAR - Int. J. Res. Anal. Rev.*, vol. 1, no. 3, pp. 2349–5138, 2014.
3. V. Kolluri, "A Comprehensive Analysis On Explainable And Ethical Machine: Demystifying Advances In Artificial Intelligence," *TIJER - Int. Res. Journals*, vol. 2, no. 7, pp. 2349–9249, 2015.

4. M. Ahmed, A. Naser Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *J. Netw. Comput. Appl.*, vol. 60, pp. 19–31, 2016, doi: https://doi.org/10.1016/j.jnca.2015.11.016.
5. V. Kolluri, "A Pioneering Approach To Forensic Insights: Utilization Ai for Cybersecurity Incident Investigations," *Int. J. Res. Anal. Rev. (IJRAR*, vol. 3, no. 3, 2016.
6. X. Huo, K. Wu, W. Miao, L. Wang, H. He, and D. Su, "Research on Network Traffic Anomaly Detection of Source-Network-Load Industrial Control System Based on GRU-OCSVM," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 300, no. 4, p. 042043, Jul. 2019, doi: 10.1088/1755-1315/300/4/042043.
7. M. S. Parwez, D. B. Rawat, and M. Garuba, "Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network," *IEEE Trans. Ind. Informatics*, vol. 13, no. 4, pp. 2058–2065, 2017.
8. P. Lin, K. Ye, and C.-Z. Xu, "Dynamic Network Anomaly Detection System by Using Deep Learning Techniques," in *Cloud Computing--CLOUD 2019: 12th International Conference, Held as Part of the Services Conference Federation, SCF 2019, San Diego, CA, USA, June 25--30, 2019, Proceedings 12*, 2019, pp. 161–176. doi: 10.1007/978-3-030-23502-4_12.
9. K. Atefi, H. Hashim, and M. Kassim, "Anomaly analysis for the classification purpose of intrusion detection system with K-nearest neighbors and deep neural network," in *Proceeding - 2019 IEEE 7th Conference on Systems, Process and Control, ICSPC 2019*, 2019. doi: 10.1109/ICSPC47137.2019.9068081.
10. I. Alrashdi, A. Alqazzaz, E. Aloufi, R. Alharthi, M. Zohdy, and H. Ming, "AD-IoT: Anomaly detection of IoT cyberattacks in smart city using machine learning," in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference, CCWC 2019*, 2019. doi: 10.1109/CCWC.2019.8666450.
11. A. Srivastava, A. Agarwal, and G. Kaur, "Novel Machine Learning Technique for Intrusion Detection in Recent Network-based Attacks," in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, IEEE, Nov. 2019, pp. 524–528. doi: 10.1109/ISCON47742.2019.9036172.
12. A. Pattawaro and C. Polprasert, "Anomaly-Based Network Intrusion Detection System through Feature Selection and Hybrid Machine Learning Technique," in *2018 16th International Conference on ICT and Knowledge Engineering (ICT&KE)*, 2018, pp. 1–6. doi: 10.1109/ICTKE.2018.8612331.
13. T. Salman, D. Bhamare, A. Erbad, R. Jain, and M. Samaka, "Machine Learning for Anomaly Detection and Categorization in Multi-Cloud Environments," in *2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud)*, IEEE, Jun. 2017, pp. 97–103. doi: 10.1109/CSCloud.2017.15.
14. O. Faker and E. Dogdu, "Intrusion Detection Using Big Data and Deep Learning Techniques," 2019. doi: 10.1145/3299815.3314439.
15. J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, Jul. 2018, doi: 10.1016/j.neucom.2017.11.077.
16. A. A. Abdulrahman and M. K. Ibrahem, "Evaluation of DDoS attacks Detection in a New Intrusion Dataset Based on Classification Algorithms," *Iraqi J. Inf. Commun. Technol.*, vol. 1, no. 3, pp. 49–55, Feb. 2019, doi: 10.31987/ijict.1.3.40.
17. S. D. D. Anton, S. Sinha, and H. Dieter Schotten, "Anomaly-based Intrusion Detection in Industrial Data with SVM and Random Forests," in *2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, IEEE, Sep. 2019, pp. 1–6. doi: 10.23919/SOFTCOM.2019.8903672.
18. Y. Zhao *et al.*, "Network Anomaly Detection by Using a Time-Decay Closed Frequent Pattern," *Information*, vol. 10, no. 8, 2019, doi: 10.3390/info10080262.
19. P. Amangele, M. J. Reed, M. Al-Naday, N. Thomos, and M. Nowak, "Hierarchical Machine Learning for IoT Anomaly Detection in SDN," *2019 Int. Conf. Inf. Technol. InfoTech 2019 - Proc.*, no. 780139, 2019, doi: 10.1109/InfoTech.2019.8860878.
20. K. Kostas, "Anomaly detection in networks using machine learning," *Res. Propos.*, vol. 23, p. 343, 2018.
21. Chinta, P. C. R., & Karaka, L. M. (2020). Agentic AI and Reinforcement Learning: Towards More Autonomous and Adaptive AI Systems.
22. Kalla, D., Kuraku, D. S., & Samaah, F. (2021). Enhancing cyber security by predicting malwares using supervised machine learning models. *International Journal of Computing and Artificial Intelligence*, *2*(2), 55-62.
23. Katari, A., & Kalla, D. (2021). Cost Optimization in Cloud-Based Financial Data Lakes: Techniques and Case Studies. *ESP Journal of Engineering & Technology Advancements (ESP-JETA)*, *1*(1), 150-157.
24. Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2021). Facial Emotion and Sentiment Detection Using Convolutional Neural Network. *Indian Journal of Artificial Intelligence Research (INDJAIR)*, *1*(1), 1-13.
25. Routhu, K., Bodepudi, V., Jha, K. M., & Chinta, P. C. R. (2020). A Deep Learning Architectures for Enhancing Cyber Security Protocols in Big Data Integrated ERP Systems. *Available at SSRN 5102662*.
26. Kuraku, S., & Kalla, D. (2020). Emotet malware—a banking credentials stealer. *Iosr J. Comput. Eng*, *22*, 31-41.
27. Kalla, D., & Samiuddin, V. (2020). Chatbot for medical treatment using NLTK Lib. *IOSR J. Comput. Eng*, *22*, 12.
28. Routhu, K., & Jha, K. M. (2021). Leveraging Big Data Analytics and Machine Learning Techniques for Sentiment Analysis of Amazon Product Reviews in Business Insights. *Available at SSRN 5106490*.
29. Chinta, P. C. R., & Katnapally, N. (2021). Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures. *Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures*.
30. Karaka, L. M. (2021). Optimising Product Enhancements Strategic Approaches to Managing Complexity. *Available at SSRN 5147875*.
31. Boppana, S. B., Moore, C. S., Bodepudi, V., Jha, K. M., Maka, S. R., & Sadaram, G. AI and ML Applications In Big Data Analytics: Transforming ERP Security Models For Modern Enterprises.