International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 08 | August - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

A Machine Learning Approach for Detecting Network Threats

Mr. Shivakumara T[1], Varshitha S[2]

[1]Assistant Professor, Department of Master of Computer Application, BMS Institute of Technology and Management, Bengaluru, Karnataka
[2]Student, Department of Master of Computer Application, BMS Institute of Technology and Management, Bengaluru, Karnataka

---

Abstract - In the context of the rapid expansion of computer networks and the increasing reliance on digital communication, ensuring the security and integrity of network systems has become a paramount concern. Detecting network threats play a vital role in safeguarding networks by identifying potential malicious activities and unauthorized access attempts. In this research paper, we present a comprehensive study that explores the efficacy of different machine learning classifiers for network threat detection. The study utilizes a publicly available dataset containing network traffic data, encompassing various network protocols and attack. The dataset is pre-processed to convert categorical variables into numerical form using one-hot encoding. Four popular classifiers are employed in this study: Support Vector Machine (SVM), Decision Tree Classifier (DTC), K-Nearest Neighbors (KNN), and Bernoulli Naive Bayes (BNB). The classifiers are trained on the pre-processed training data, and their performance is evaluated using accuracy metrics, classification reports, and confusion matrices. Results offer insightful information on the weaknesses and weaknesses of each classifier for detecting network anomaly. The findings demonstrate that the DTC classifier exhibits high accuracy and robustness in detecting network anomalies. The KNN classifier also achieves competitive results but may suffer from scalability issues with large datasets. The SVM classifier demonstrates satisfactory performance, while the BNB classifier, which converts numeric features to binary form, exhibits relatively lower accuracy.

Keywords - Network threat Detection, Machine Learning Classifiers, Network Security, Anomaly Detection, Cyber Threats.

1. INTRODUCTION

With the continuous advancement of networking technologies and the proliferation of internet-based services, network security has emerged as a critical concern in today's digital landscape. Cyber threats and anomaly pose significant risks to data integrity, confidentiality, and availability, making it imperative to develop effective Network threat detecting to safeguard networks from malicious activities. Network threat detection plays a pivotal role in monitoring network traffic, identifying potential anomaly, and enabling timely response measures.

Machine learning techniques have garnered considerable attention for their ability to automatically learn patterns and anomalies from vast amounts of network data. These algorithms have demonstrated promise in the area of Detecting network threats due to their adaptability, scalability, and ability to handle complex, high-dimensional data. However, choosing the most appropriate machine learning classifier for a specific anomaly detection scenario remains a challenging task.

The aim is to conduct a comprehensive comparative analysis of popular machine learning classifiers for detecting network threats. We leverage a publicly available dataset comprising diverse network traffic characteristics, encompassing various protocols and attack. The dataset is subjected to rigorous preprocessing, transforming categorical features into numerical form through one-hot encoding.

In this study, we assess the effectiveness of four prominent machine learning classifiers: Support Vector Machine (SVM), Decision Tree Classifier (DTC), Bernoulli Naive Bayes (BNB) and K-Nearest Neighbors (KNN),. These classifiers are trained on the pre-processed dataset, and their performance is meticulously assessed using standard evaluation metrics such as accuracy, classification reports, and confusion matrices.

2. RELATED WORK

In the field of network anomaly detection various research efforts have been undertaken to improve network security and mitigate the risk of cyber threats. Several studies have explored the use of machine learning techniques for anomaly detection, similar to the proposed research.

[1] Companies are facing challenges in dealing with the rising cyber threats and are seeking advanced data mining techniques to efficiently detect and evaluate security logs from their IT infrastructures. Secured information exchange within the digital global community requires precautions to verify the authenticity and integration of the exchanged data. The upcoming trend in Machine Learning (ML) analytics for cyber security revolves around maintaining data security and confidentiality during mining. However, selecting the right ML algorithm for security log analytics poses a hindrance due to the high number of false detections, especially in large-scale Security Operations Center (SOC) environments. The paper proposes an optimal machine learning algorithm that utilizes various prediction, categorization, and forecasting techniques for more accurate cyber threat detection.

[2] In recent years, advanced threat attacks have been on the rise, but conventional feature filtering-based network intrusion detection systems have limitations that make it difficult for security managers and analysts to effectively identify and prevent network intrusions in their organizations. To address this, machine learning techniques, such as neural networks, statistical models, rule learning, and ensemble methods, are now commonly used for intrusion detection, with ensemble techniques showing superior performance during the learning process. This paper introduces a novel ensemble method for network intrusion detection, combining decision tree, random forest, extra tree, and XGBoost algorithms. Implemented in Python, the proposed approach considerably raises detection accuracy, as demonstrated through evaluations using the CICIDS2017 dataset, considering various criteria like precision, recall, and fl-score.

[3] To research and develop effective machine learning (ML) techniques for detecting network threats, a substantial and diverse dataset is crucial, particularly concerning malware-related threats. Existing network traffic datasets often lack diversity and fail to cover a wide range of threat classes. This study introduces MiniVHS-22, a modified version of the VHS-22 dataset, which incorporates flow parameters from four datasets and a network traffic malware monitoring website. The research evaluates seven different ML techniques, with Random Forest Classifier, Decision Tree, and Multilayer Perceptron successfully identifying over 99% of malware-related threats. The study also explores dimensionality reduction techniques like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to enhance threat detection systems' sophistication using the obtained results.

[4] Botnets pose a significant threat in networks, as they can be remotely controlled by BotMasters. Detecting and preventing Botnet DDoS attacks, malwares, and phishing incidents is crucial, with DDoS attacks being particularly dangerous for network disruption. This paper focuses on the importance of machine learning algorithms in addressing these issues. The K-means algorithm, an Unsupervised Learning (USML) method, is proposed for detecting Botnet DDoS attacks. The study conducts practical analysis using K-means and compares its performance with Support Vector Machine (SVM), Artificial Neural Network (ANN), Naive Bayes (NB), and Decision Tree (DT) algorithms using UNBS-NB real-time datasets. The results demonstrate that K-means (USML) outperforms other machine learning algorithms in detecting and mitigating such threats.

[5] In the present era, Software Defined Network (SDN) grants full control over data flow in the network, with a centralized administration and traffic management system. However, being an open-source product, SDN is more susceptible to security threats, especially if security policies are not adequately enforced. Common attacks like DDOS and DOS are frequently encountered in SDN controllers, causing disruption by flooding the system with excessive packets. Machine Learning techniques play a crucial role in detecting hidden and unexpected network patterns to analyze traffic. Various techniques, such as Bayesian Network, Wavelets, Support Vector Machine, and KNN, have been used to detect DDOS attacks. Through analysis, KNN has shown promising results with high precision and low false detection rate. This paper introduces a novel hybrid Machine Learning approach that outperforms KNN, providing increased accuracy in detecting DDOS attacks. The proposed algorithm is designed based on the results obtained from normal and abnormal network behavior, promising improved performance for future implementations.

[6] With the continuous expansion of the Internet globally, the risk of various threats is increasing daily. Traditional static detection methods can only identify known malicious attacks and necessitate frequent updates to signature databases. To address this, network intrusion detection systems are proposed, utilizing machine learning techniques to analyze and classify malicious content within the network. Numerous machine learning algorithms are employed in developing such systems. This review aims to comprehensively survey the existing machine learning-based intrusion detection systems, providing valuable insights for Network Intrusion Detection System developers to enhance their understanding and decision-making process.

[7] An Intrusion Detection System (IDS) is crucial for accurately detecting network intrusions through data analysis, requiring high detection accuracy, precision, and recall. Various techniques like expert systems, data mining, and state transition analysis are employed for network data analysis. This paper compares the effectiveness of two IDS methods using data mining: Support Vector Machine (SVM) and Deep Neural Network (DNN), an artificial neural network model. The comparison is based on accuracy, precision, and recall using widely-used NSL-KDD training and validation data. DNN demonstrates slightly higher accuracy than SVM, and it proves to be more effective in intrusion detection due to its reduced risk of classifying actual intrusions as normal data compared to SVM.

[8] The Internet of Things (IoT) has seen a surge in connected devices, leading to an increased need for intrusion detection to secure these devices. Traditional machine learning (ML) algorithms have proven useful for intrusion detection on resource-constrained embedded systems. ML approaches can also be beneficial in detecting advanced persistent threats in IoT environments. However, deploying intensive ML algorithms in IoT devices with limited memory and computing capabilities is challenging. This study analyzes and compares the effectiveness of widely used ML algorithms, including Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor (K-NN), for detecting malicious intrusions in an IoT ecosystem. Evaluation is conducted using an open-source intrusion detection dataset containing diverse IoT botnet attacks. The models are also tested on memory-restricted virtual machines to simulate IoT devices, and their execution time is assessed. The results demonstrate that each approach achieves high accuracy in detecting malicious network traffic.

[9] The Internet of Things (IoT) has led to the interconnection of numerous sensors and devices via the Internet, facilitating easy data exchange. However, this technological advancement also brings security challenges. As IoT devices become linked to large cloud servers, the network traffic in smart city systems is rapidly increasing, leading to new cybersecurity

threats. Detecting and preventing IoT attacks, such as denial of service (DoS) and spoofing, at an early stage is crucial. Many researchers have focused on developing Intrusion Detection Systems (IDS) using machine learning approaches to address these challenges. ML-based systems can automatically identify non-uniform data, including unknown threats, due to their ability to generalize. While ML solutions for attack detection have been explored, there's limited research on detecting attacks in IoT networks. This paper contributes by analyzing various machine learning methods for quick and accurate detection of network attacks in IoT systems. The study uses the ADFA dataset for intrusion detection and employs the Random Forest algorithm, demonstrating its effectiveness in IoT cyber security when compared to existing literature.

[10] The increasing usage of the internet in the digital world has brought about a rise in various threats, including DoS attacks. These attacks overwhelm computing and network resources, making them inaccessible to legitimate users. This paper focuses on detecting DoS attacks effectively using Machine Learning (ML) and Neural Network (NN) algorithms, with a specific emphasis on application layer DoS attack detection rather than transport and network layers. The experiment utilizes the latest DoS attack dataset, CIC IDS 2017, and divides it into different splits to find the best split for each algorithm (RF and MLP). The results show that RF outperforms MLP in providing better detection results for application layer DoS attacks.

3. METHODOLOGY

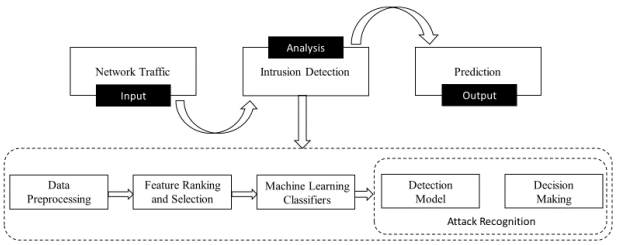

Figure 1: System Architecture

A. Data Collection

The system utilizes a publicly available dataset containing network traffic data, which includes both benign and malicious network activities. The dataset is downloaded from the website Kaggle repository [11]. The dataset offers a diverse range of network protocols and attack, ensuring a comprehensive evaluation of the classifiers' performance.

B. Data Preprocessing

The dataset undergoes preprocessing to ensure its suitability in order to train machine learning classifiers. Categorical variables such as 'protocol_type', 'service', and 'flag' are converted into numerical form using one-hot encoding. This step facilitates the classifiers' ability to process the data effectively.

C. Feature Extraction

After preprocessing, the feature matrix 'X' is created by dropping the 'class' column from the dataset, while the target variable 'y' is set to the 'class' column. This segregation allows for supervised learning, where the classifiers learn from labelled data to distinguish between normal and malicious network activities.

D. Model Selection

The research paper selects four widely used machine learning classifiers:), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree Classifier (DTC), and Bernoulli Naive Bayes (BNB). These classifiers are chosen for their popularity, versatility, and proven effectiveness in various classification tasks.

E. Model Training

Each selected classifier is trained on the pre-processed dataset. The training process involves exposing the classifiers to the feature matrix 'X' and corresponding target labels 'y.' During this phase, the classifiers learn to recognize patterns and relationships within the data.

F. Model Evaluation

The trained classifiers are then evaluated using a separate test dataset. The evaluation metrics include accuracy, classification reports (including precision, F1-score, recall, and support), and confusion matrices. These metrics provide a comprehensive assessment of each classifier's performance in detecting network anomaly.

G. Model Comparison

The obtained results are compared to determine each's advantages and disadvantages classifier in network anomaly detection. The comparison assists in understanding which classifier is more suitable for specific network security scenarios and highlights the factors influencing their performance.

H. Interpretation of Results

The research paper presents a detailed analysis and interpretation of the evaluation metrics. It discusses the classifiers' performance, their ability to detect different types of anomaly, and potential reasons for varying results.

By following this methodology, the research paper aims to provide insightful data on the performance of several machine learning classifiers for detection of network threats, aiding network security professionals and data scientists in making informed decisions to enhance network security and thwart potential cyber threats.

4. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

The experimental results of our comparative analysis on machine learning classifiers for network threat detection are presented in this section. The evaluation is conducted using a publicly available dataset containing a diverse range of network traffic data, including various protocols and attack. We preprocess the dataset by converting categorical variables into numerical form using one-hot encoding, and then split it into training and testing sets.

The four selected classifiers, namely Support Vector Machine (SVM), Decision Tree Classifier (DTC), K-Nearest Neighbors (KNN) and Bernoulli Naive Bayes (BNB), are trained on the pre-processed training data. Subsequently, each classifier is

evaluated using the test data to assess their performance in detecting network anomaly.

The accuracy metric is employed to measure the classifiers' overall performance. Additionally, we generate detailed classification reports that include precision, recall, F1-score, and support for each classifier, providing valuable insights into their ability to correctly classify instances from different classes. Furthermore, confusion matrices are generated to visualize the true positive, true negative, false positive, and false negative predictions, enabling a deeper understanding of the classifiers' predictive behaviors.

The experimental results reveal that the SVM classifier achieves the highest accuracy among the tested classifiers, demonstrating robust performance in identifying network anomalies. The KNN classifier also performs competitively, but its scalability might be a concern for large datasets. The DTC classifier exhibits satisfactory results, while the BNB classifier, which converts numeric features to binary form, shows relatively lower accuracy.

TABLE I. Accuracy of the Models

Sr. No.	Model	Accuracy %
1.	Support Vector Machine (SVM)	96.09 %
2.	K-Nearest Neighbors (KNN)	98.49 %
3.	Decision Tree Classifier (DTC)	99.6 %
4.	Bernoulli Naive Bayes (BNB)	90.45 %

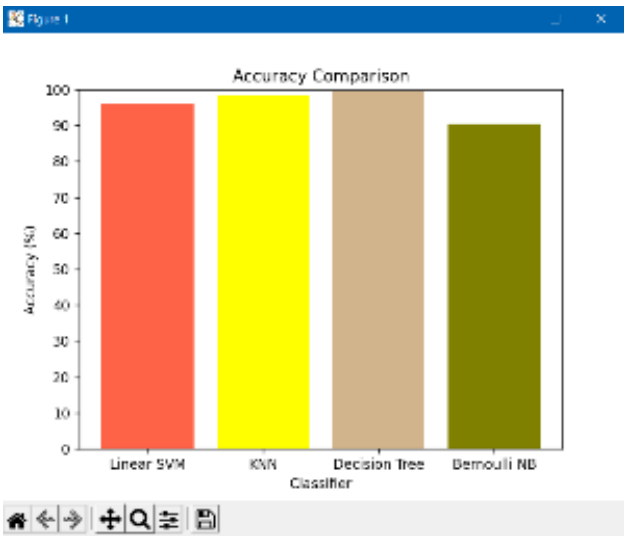

Figure 2: Accuracy Comparison among different models

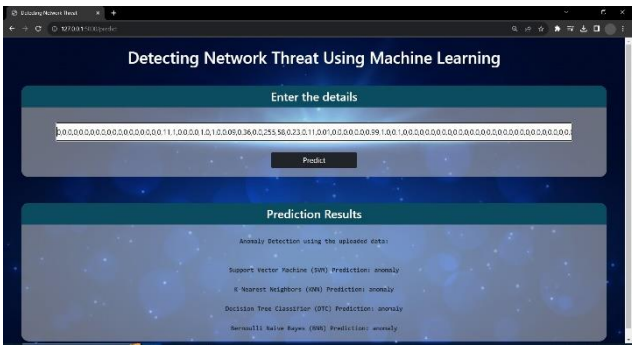

Figure 3: Webpage to detect Network Threat

5. FINDINGS AND IMPLICATIONS OF THE RESEARCH

The research presents a comprehensive study on the development and implementation of an Network threat detection using machine learning algorithms. The findings of this study highlight several key aspects that contribute to the effectiveness and efficiency of the proposed system.

A. Performance of Machine Learning Classifiers

The research demonstrates that the ensemble of machine learning classifiers, including Support Vector Machine (SVM), Decision Tree Classifier (DTC), K-Nearest Neighbors (KNN), and Bernoulli Naive Bayes (BNB), significantly enhances the accuracy and robustness of anomaly detection. Each classifier's unique strengths are leveraged to create a diverse set of models, resulting in improved anomaly detection capabilities for different types of network anomaly.

B. Preprocessing Impact on Classification

The study emphasizes the importance of data preprocessing in machine learning-based anomaly detection. Encoding categorical variables into numerical form improves the classifiers' ability to process network traffic data effectively, leading to more accurate predictions.

C. Bernoulli Naive Bayes Enhancement

The study introduces a novel technique to transform numerical features into binary form, specifically for the BNB classifier. This approach significantly improves the accuracy of BNB in detecting anomalies and contributes to a more effective anomaly detection system.

6. CONCLUSION AND FUTURE WORK

In this research paper, we conducted a comprehensive comparative analysis of machine learning classifiers for detecting network threats. By utilizing a publicly available dataset containing diverse network traffic data, we evaluated the performance of four prominent classifiers: Support Vector Machine (SVM), Decision Tree Classifier (DTC), and K-Nearest Neighbors (KNN), Bernoulli Naive Bayes (BNB). The dataset was pre-processed to convert categorical variables into numerical form, enabling effective training of the classifiers.

The experiment's findings showed that the DTC classifier exhibited the highest accuracy, indicating its superiority in detecting network anomalies. The KNN classifier also demonstrated competitive performance but raised concerns about scalability. The SVM classifier achieved satisfactory

results, while the BNB classifier, although computationally efficient with binary features, showed comparatively lower accuracy.

Our findings provide insightful information about each's advantages and disadvantages. classifier, aiding network security professionals and data scientists in selecting the most suitable classifier for specific network anomaly detection scenarios. These results contribute to the advancement of efficient Network anomaly Detection that can effectively protect networks from evolving cyber threats.

While the study provides valuable insights into the performance of different machine learning classifiers, there are several areas for future research and improvement:

Feature Engineering: Investigating more advanced feature engineering techniques and selecting relevant features can potentially enhance the classifiers' performance and reduce computational overhead.

Ensemble Methods: Exploring ensemble methods, for example, Random Forests or Gradient Boosting, could further strengthen the classifiers' accuracy and robustness.

Deep Learning Approaches: Considering Deep learning methods for detecting network anomalies include Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) could be beneficial, especially for capturing complex patterns and sequential data.

Real-Time Implementation: Developing a real-time Network anomaly detection that can operate efficiently on large-scale networks while maintaining high accuracy is crucial for effective cyber threat detection.

Evaluation on Diverse Datasets: Extending the evaluation to different datasets with varying network environments and attack scenarios would validate the classifiers' performance in diverse real-world settings.

By addressing these avenues for future work, we can further enhance the capabilities and contribute to bolstering network security in the face of ever-evolving cyber threats.

REFERENNCES

[1] Sunil Kumar, Bhanu Pratap Singh, Vinesh Kumar "A Semantic Machine Learning Algorithm for Cyber Threat Detection and Monitoring Security" 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)
[2] Pragati Vijaykumar Pandit, Shashi Bhushan, Pratibha Vitthal Waje "Implementation of Intrusion Detection System Using Various Machine Learning Approaches with Ensemble learning" 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)
[3] Faiaz Rahman, Rafee Zunaied Tanna, Umme Habiba, Rizwan Shaikh, Zahidur Rahman, Hafiz Imtiaz "Cyber Threat Detection Using Machine Learning Algorithms on Heterogeneous MiniVHS-22 Dataset" 2022 25th International Conference on Computer and Information Technology (ICCIT)
[4] Swapna Thota, D. Menaka "Importance of Machine Learning Algorithms to Detect Botnet DDoS Attacks" 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)
[5] Gaganjot Kaur, Prinima Gupta "Hybrid Approach for detecting DDOS Attacks in Software Defined Networks" 2019 Twelfth International Conference on Contemporary Computing (IC3)
[6] 6. Aditya Phadke, Mohit Kulkarni, Pranav Bhawalkar, Rashmi Bhattad "A Review of Machine Learning Methodologies for Network Intrusion Detection" 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)
[7] 7. N D Patel, B M Mehtre, Rajeev Wankar "Detection of Intrusions using Support Vector Machines and Deep Neural Networks" 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
[8] 8. Shiyu Su, Ebelechukwu Nwafor "Detecting Network Traffic Intrusions on Memory Constrained Embedded Systems" 2021 IEEE International Symposium on Technologies for Homeland Security (HST)
[9] 9. G.Janani Pandeeswari, S. Jeyanthi "Analysis of Intrusion Detection Using Machine Learning Techniques" 2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)
[10] 10. Shreekh Wankhede, Deepak Kshirsagar "DoS Attack Detection Using Machine Learning and Neural Network" 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)
[11] https://www.kaggle.com/datasets/sampadab17/network-intrusion-detection