IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 6, June 2023

DOI: 10.17148/IARJSET.2023.10668

Network Anomaly Detection Using Machine Learning: A Comprehensive Study

Narla Sai Vardhan Reddy[1], **Koppula Arun**[2], **Manideep Mallurwar**[3],

Panthangi Venkateswara Rao[4]

Student, Information Technology, Mahatma Gandhi Institute of Technology, Hyderabad, India[1]

Student, Information Technology, Mahatma Gandhi Institute of Technology, Hyderabad, India[2]

Student, Information Technology, Mahatma Gandhi Institute of Technology, Hyderabad, India[3]

Assistant Professor, Information Technology, Mahatma Gandhi Institute of Technology, Hyderabad, India[4]

Abstract: Millions of individuals use the Internet daily to communicate with thousands of enterprises. Attacks on networks that have never occurred may come from abnormalities. Even though it has been researched for a long time, it is always difficult to identify and guard networks against unapproved access. Alongside these improvements, there are an increasing number of everyday Internet attacks. The proposed system uses machine learning techniques to find network anomalies. The CICIDS2017 dataset was utilised to accomplish this because of how current and diverse the attacks are. To determine accuracy, precision, recall, and f1-score, various machine learning techniques, including Naive Bayes, Random Forest, ID3, K Nearest Neighbours, and MLP, were utilised.

Keywords: CICIDS2017, Naive Bayes, Random Forest, ID3, K Nearest Neighbours, MLP.

I. INTRODUCTION

Organizations encounter increasing challenges in maintaining the safety of their private information in today's fast-growing networking environment. An increase in unidentified attacks has been observed from inside and outside sources. Two main techniques are used to identify attacks to maintain the security of information: based on signatures classification and detection based on anomalies. A database containing known signatures of attacks is used by signature-based techniques to identify threats. This strategy has worked well, but it necessitates ongoing database upgrades and analysis of new attack data.

Furthermore, even with up-to-date databases, signature-based techniques are vulnerable to newly discovered zero-day attacks. Such attacks can get past the detection system even though they are not included in the database. Anomaly-based techniques, on the other hand, concentrate on identifying unexpected network behaviours through network flow analysis. Devices for network monitoring collect a lot of data quickly.

II. RELATED WORKS

Anomaly detection has been the subject of numerous studies, demonstrating the importance of this field of study across numerous industries. To give readers a thorough overview of the topic, we will examine the body of existing knowledge and developments in anomaly detection approaches in this literature survey.

The significance of anomaly-based network intrusion detection systems for safeguarding networks from hostile activity is highlighted by Veeramreddy and Prasad [1]. They place a strong emphasis on the use of machine learning and data mining methods in the development of efficient intrusion detection systems. The authors discuss the shortcomings of classification-based methods for identifying unidentified attacks and advise looking into unsupervised learning techniques to understand dynamic intrusion operations better.

They suggest using a generalised meta-heuristic scale that incorporates canonical correlation analysis with feature optimisation techniques to attain high detection rates and low false alarm rates. Experimental findings on the NSL-KDD dataset show that their solution is superior to existing ones regarding accuracy and performance measures.

A brand-new method for computer network intrusion detection termed the Outlier Detection technique is presented by J. Jabez and B. Muthukumar [2]. Their work focuses on the effective identification of anomalies in the network activity,

DOI: 10.17148/IARJSET.2023.10668

utilising the Neighbourhood Outlier Factor (NOF) to quantify the dataset of anomalies. To improve the performance of the Intrusion Detection System (IDS) 's performance, the suggested technique uses a trained model with large datasets in a distributed storage environment. Comparing the suggested technique to other machine learning methods currently in use, experimental findings show the proposed methodology to be more effective at discovering anomalies. The study emphasises the significance of early assault detection to lessen the impact of attacks. The study helps to increase the effectiveness of IDS and can be expanded even further to include distance computation functions between the trained model and the testing model.

A survey on anomaly detection in Network Intrusion Detection Systems (IDS) utilising Particle Swarm Optimisation (PSO)-based machine learning approaches is presented by Satpute et al. in [3]. They draw attention to the drawbacks of conventional security measures, the difficulties of real-time detection, and the identification of fresh attacks. The research investigates using PSO in conjunction with several machine learning methods, such as Support Vector Machines and Neural Networks, to improve IDS performance. The authors integrate PSO and machine learning approaches to enhance anomaly detection in IDS, emphasising the necessity for a holistic approach to network security.

The difficulties of anomaly detection in Network Intrusion Detection Systems (NIDSs) are discussed by Zhang and Zulkernine [4].

A method employing unsupervised outlier identification with the random forest algorithm is suggested. Traditional NIDSs based on trained algorithms experience large false positive rates due to a lack of attack-free training data and shifting network settings. By recognising outliers, the proposed framework uses random forests to create patterns of network services and detect intrusions. The authors use the KDD'99 dataset to assess their method and provide updates to the outlier detection system. The outcomes show how well the suggested unsupervised anomaly detection method is.

A. Naive Bayes

The Bayesian classification Naive Bayes is frequently employed in anomaly identification. It is effective for high-dimensional data since it implies that the characteristics are dependent upon being given a class label. Utilising the attribute's likelihoods, Naive Bayes estimates the likelihood that a particular event is abnormal [5].

B. K Nearest Neighbours

A non-parametric supervised learning technique for classification and regression is the k-nearest neighbours (k-NN) algorithm [6]. It uses a dataset's k closest training samples to decide the output. Performance depends heavily on k, distance metric, and feature scaling. The local data structure is important for k-NN because noisy features can have an impact. It is a versatile and well-liked machine learning algorithm.

C. ID3

The ID3 decision tree technique uses information gain for attribute selection, but it has issues processing attribute values and determining how important an attribute is. To solve these problems, a novel method for attribute weighting based on simulated conditional probability is given, considering the attributes' closeness to the decision attribute.

The weighted qualities and information gain are coupled to increase the decision outcomes' correctness. According to experimental findings, the revised algorithm works better than the standard ID3 method regarding predicted accuracy and leaf count [7].

D. Random Forest

A well-liked machine learning approach called random forest combines different decision trees to get reliable results [8]. It produces more accurate predictions by utilising bagging and feature randomisation to combat overfitting. The versatility of random forest makes it possible to easily determine the value of a feature while performing both classification and regression tasks. It also has uses in e-commerce, healthcare, and other fields [8].

E. MLP

Anomaly detection frequently uses a particular kind of artificial neural network called MLP. It has many layers of linked neural networks and can identify complicated connections in the data. MLP algorithms can detect irregular correlations among parameters and undergo training using backpropagation. It can be used to find abnormalities in unstructured and structured information; however, compared to other algorithms, it can need more training data and computer power.

DOI: 10.17148/IARJSET.2023.10668

III. PROPOSED SYSTEM

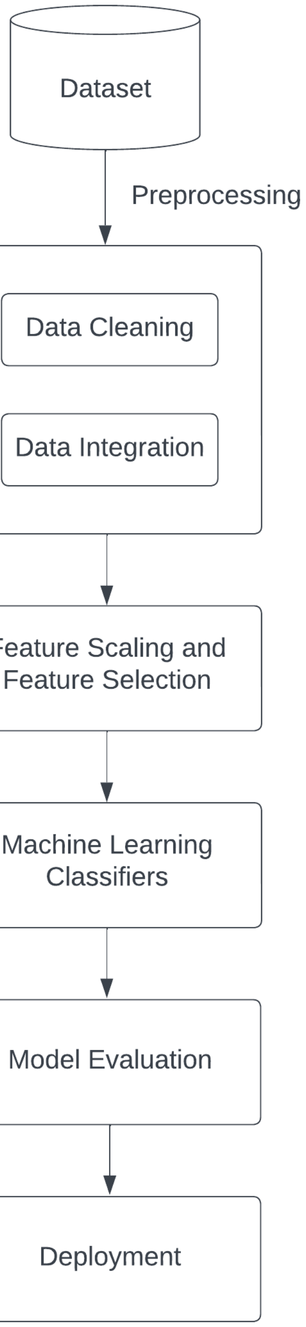

Fig. 1 Flow of Proposed System

A. Data Preprocessing

In the proposed solution to address the limitations of the existing system, the CICIDS 2017 dataset is utilized [9]. This dataset is employed to handle issues such as noise, missing values, and incompatible formats. The study combines eight CSV files from the dataset to create a unified dataset for analysis.

B. Attack Filtration

Attack filtration concentrates on separating the dataset's 12 attack kinds. A balanced distribution of attack and benign instances separates each attack into its CSV file.

C. Attribute Selection

Choosing the variables that will be used to build the models. We identify the main characteristics that differentiate between assaults and legitimate traffic by examining the attack files.

D. Implementation of machine learning

The train-test split process is used to assess the effectiveness of models. The models are tested on the test dataset after being evaluated on the training data. The chosen features from the feature selection module are used to generate performance metrics, including accuracy, precision, recall, and f1-score.

E. Deployment

It entails developing a user interface where users can enter values for seven attributes about network traffic or system behavior. The input is processed by the system, which then determines whether or not it indicates an anomaly. Users can view the outcome on a specific page, which gives them access to real-time information on the existence of anomalies.

IV. RESULTS

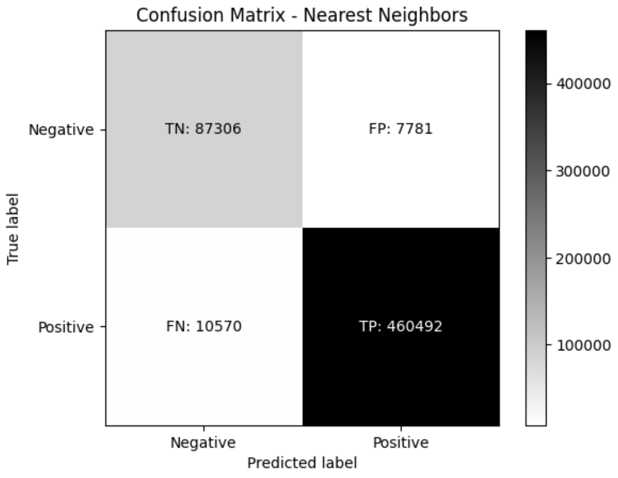

Fig. 2 Confusion matrix of KNN

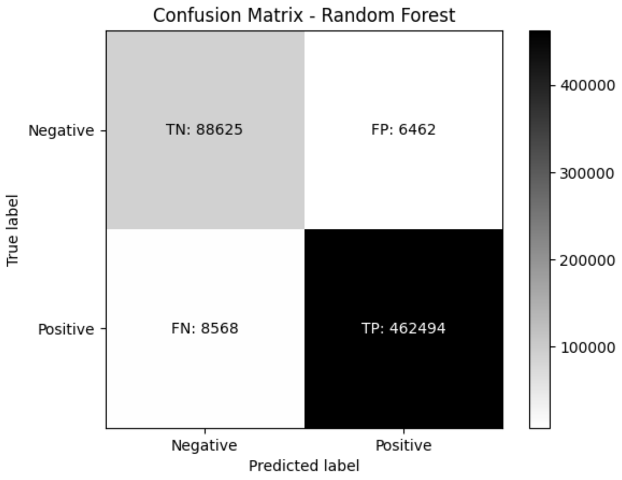

Fig. 3 Confusion matrix of Random Forest

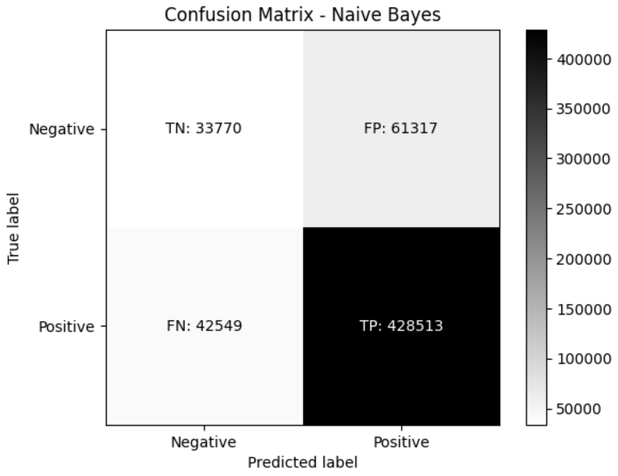

Fig. 4 Confusion matrix of Naive Bayes

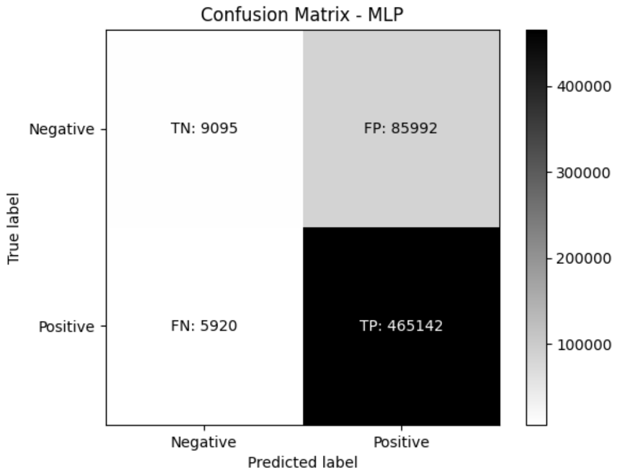

Fig. 5 Confusion matrix of MLP

DOI: 10.17148/IARJSET.2023.10668

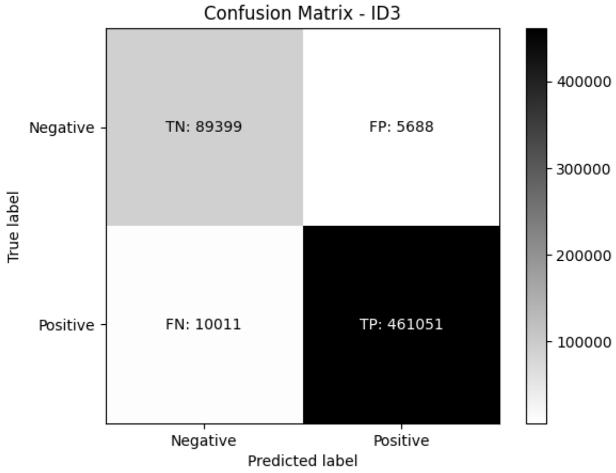

Fig. 6 Confusion matrix of ID3

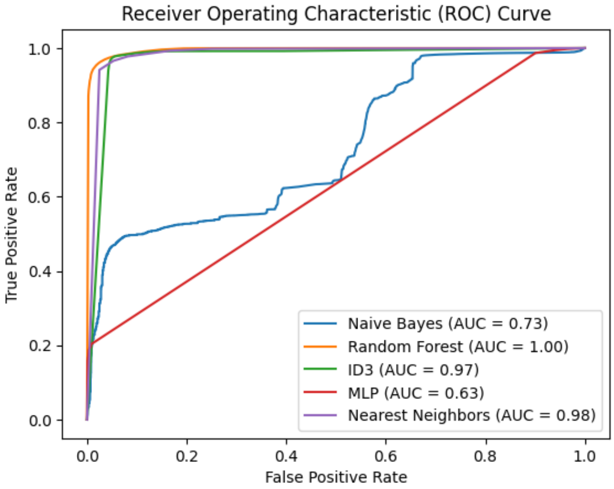

Fig. 7 ROC Curve for all five algorithms

The visualisations make it clear that Naive Bayes has a higher percentage of false positives than other algorithms, which leads to lower accuracy. On the other hand, the accuracy of Random Forest and ID3 is higher because of their balanced performances and comparatively low rates of false positives and false negatives. MLP displays many true positives despite having a higher proportion of false positives, demonstrating its accuracy in identifying positive cases. Nearest Neighbour performs well with few false positives and false negatives, improving accuracy. The selection of an algorithm should be based on the particular needs and priorities of the current classification task.

TABLE I PERFORMANCE MEASURES FOR FIVE ALGORITHMS

Classifier	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.82	0.88	0.91	0.89
Random Forest	0.97	0.94	1.00	0.97
ID3	0.97	0.96	0.98	0.97
MLP	0.84	0.85	0.99	0.91
Nearest Neighbors	0.97	0.98	0.98	0.98

Based on the exceptional performance metrics observed, it is evident that the algorithms have demonstrated remarkable capabilities in terms of accuracy, precision, recall, and F1-score. Among them, Random Forest has exhibited outstanding performance with an impressive accuracy of 0.97, precision of 0.94, recall of 1.00, and an exceptional F1-score of 0.97. These remarkable results unequivocally establish Random Forest as the most superior algorithm for the classification task at hand.

V. CONCLUSION

In this research, abnormalities in networks are discovered using machine learning approaches. Because of its currentness, vast attack inclusion, and range of network protocols, the CICIDS2017 dataset is being employed in this situation. More than 80 features define the network flow in this dataset. The application used the Random Forest Regressor algorithm to determine their significance weights and choose the features used in machine learning approaches. Before determining relevance weights for this category, the suggested strategy classifies every assault into the same category.

After conducting thorough research, we employed five diverse machine learning algorithms, each with unique attributes. Following rigorous performance evaluation, we unequivocally determined that Random Forest displayed unparalleled proficiency in accuracy, precision, recall, and F1-score. As a result, we confidently selected Random Forest as the ultimate choice for the final model deployment. This research contributes to the advancement of anomaly detection in network security and highlights the effectiveness of Random Forest in addressing such challenges.

DOI: 10.17148/IARJSET.2023.10668

REFERENCES

[1]. Veeramreddy, Jyothsna, & Prasad, Koneti. (2019). "Anomaly-Based Intrusion Detection System." In Proceedings of the book "Intrusion Detection Systems". IntechOpen. DOI: 10.5772/intechopen.82287.

[2]. Jabez, J., & Muthukumar, B. (2015). "Intrusion Detection System (IDS): Anomaly Detection Using Outlier Detection Approach." Procedia Computer Science, 48, 338-346. ISSN 1877-0509. DOI: 10.1016/j.procs.2015.04.191.

[3]. Satpute, K., Agrawal, S., Agrawal, J., Sharma, S. (2013). "A Survey on Anomaly Detection in Network Intrusion Detection System Using Particle Swarm Optimization Based Machine Learning Techniques." In Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA). Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-35314-7_50.

[4]. Zhang, J., & Zulkernine, M. (2006). "Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection." In 2006 IEEE International Conference on Communications (pp. 2388-2393). Istanbul, Turkey. DOI: 10.1109/ICC.2006.255127.

[5]. John, George H., and Pat Langley. (1995). “Estimating continuous distributions in Bayesian classifiers.” In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence (pp. 338-345). Morgan Kaufmann Publishers Inc.

[6]. Wikipedia contributors. (2023, June 18). K-nearest neighbors algorithm. In Wikipedia, The Free Encyclopedia. Retrieved 13:27, June 23, 2023, from https://en.wikipedia.org/w/index.php?title=K-nearest_neighbors_algorithm&oldid=1160701344

[7]. Liang, X., Qu, F., Yang, Y., & Cai, H. (2015). An Improved ID3 Decision Tree Algorithm Based on Attribute Weighted. Retrieved from https://www.atlantis-press.com/article/25841452

[8]. IBM. "What is random forest?" [Online]. Available: https://www.ibm.com/topics/random-forest

[9]. Sharafaldin, I., Lashkari, A.H., Ghorbani, A.A. (2018). Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, January 2018.