IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

Anomaly Detection in Cybersecurity Using Random Forest

Mr. Ashish Modi[1] and Mr. Kunj Navadiya[2]
Asst. Prof, Department of Computer and Information Science[1]
Student, Department of Computer and Information Science[2]
Nagindas Khandwala College, Mumbai, Maharashtra, India

Abstract: *Cybersecurity threats are increasing at an alarming rate, making effective detection methods more important than ever. Traditional security systems often struggle to keep up with evolving threats, which is where machine learning comes into play. Random Forest (RF), a powerful ensemble learning technique, has proven to be highly effective for anomaly detection. This paper explores how RF can be applied to cybersecurity, highlighting its strengths, practical implementation, and how it compares to other approaches. Our findings, using the UNSW-NB15 dataset, show that RF can identify malicious activities, particularly DoS and Backdoor attacks, with impressive accuracy while keeping false positives low. Additionally, we compare RF with deep learning-based techniques and discuss future improvements for real-time cybersecurity applications.*

Keywords: Cybersecurity, Anomaly Detection, Machine Learning, Random Forest, Network Security.

I. INTRODUCTION

With the rapid expansion of digital infrastructure, cybersecurity threats have become more pervasive, posing a significant challenge to organizations and individuals alike. Cyber-attacks such as malware injections, phishing, denial-of-service (DoS), and unauthorized access continue to evolve in complexity. Traditional signature-based and heuristic security measures often fail to detect sophisticated threats, necessitating the use of data-driven techniques such as machine learning.

Among various machine learning techniques, Random Forest (RF) has gained popularity for anomaly detection due to its ensemble-based approach, high accuracy, and interpretability. RF is particularly effective in identifying patterns in network traffic, user behavior, and system logs, making it a valuable tool in cybersecurity.

This paper provides a comprehensive study on the application of RF for anomaly detection in cybersecurity, including its advantages, implementation, and performance comparison with deep learning models. The contributions of this study are:

1. Analyzing the efficacy of RF for anomaly detection using the UNSW-NB15 dataset.
2. Evaluating the model's performance using various metrics such as accuracy, precision, recall, and F1-score.
3. Comparing RF with deep learning models, highlighting their trade-offs.
4. Discussing future research directions to enhance cybersecurity anomaly detection.

II. BACKGROUND AND LITERATURE REVIEW

Cybersecurity anomaly detection has been a focal area of research, with various methodologies being proposed over the years. Traditional rule-based intrusion detection systems (IDS) have been widely used but often fail to detect zero-day attacks due to their reliance on predefined attack signatures.

Traditional Approaches to Anomaly Detection

Traditional anomaly detection techniques include:

- **Signature-Based Detection:** Uses predefined signatures to detect threats, but struggles against new attack patterns.
- **Rule-Based Systems:** Employ predefined rules but require constant updates.
- **Statistical Methods:** Identify deviations from normal behavior but often generate high false positives.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

Machine Learning in Cybersecurity

Machine learning models such as Support Vector Machines (SVM), Decision Trees (DT), and Neural Networks have been explored for anomaly detection. RF, an ensemble-based learning technique, has been favored due to its ability to handle high-dimensional data, robustness to noise, and interpretability.

Several studies have demonstrated the efficiency of RF in cybersecurity applications. For instance, [Research Reference] compared RF with SVM and deep learning models, showing that RF performs competitively while being computationally efficient. Our research builds on these findings, extending them by incorporating a comprehensive evaluation of RF against deep learning models.

II. LITERATURE REVIEW

Moustafa and Slay (2015) created the UNSW-NB15 dataset to address the shortcomings of earlier datasets like KDD99. This new dataset offers a more realistic snapshot of network traffic by including modern attack techniques and a wider range of network features. Their work highlights how important it is to use high-quality datasets when building machine learning models for intrusion detection, ultimately helping to improve the overall performance of these systems.

In 2016, Moustafa and Slay took a closer look at both the UNSW-NB15 and KDD99 datasets. By comparing their statistical properties, they assessed how well each one performs in training anomaly detection models. Their research underscores how crucial it is to work with diverse and realistic datasets to ensure that machine learning models generalize well, particularly in the complex world of network intrusion detection.

In 2017, Moustafa and his team introduced a new technique: Geometric Area Analysis (GAA). This method uses trapezoidal area estimation to spot anomalies in large-scale networks. By bringing geometry into the mix, the team improved detection accuracy, helping systems better distinguish between normal and malicious network behavior. This approach is a fresh take on anomaly detection, adding depth to existing methodologies.

Later that same year, Moustafa et al. explored how big data analytics could be applied to intrusion detection. They proposed a statistical decision-making model that leverages Finite Dirichlet Mixture Models (FDMMs) to analyze network traffic. Their work shines a light on how probabilistic modeling can significantly boost the accuracy and effectiveness of intrusion detection systems, especially in large-scale environments where volume and complexity are key challenges.

Fast forward to 2020, and Sarhan et al. introduced another breakthrough: NetFlow-based datasets for machine learning in intrusion detection. Recognizing the need for scalable datasets that reflect the ever-evolving nature of cyber threats, they designed datasets that incorporate NetFlow features. Their research provides new insights into how we can build more robust and adaptable intrusion detection systems that are better equipped to handle today's cybersecurity challenges.

III. METHODOLOGY

This section outlines the methodology used to implement and evaluate Random Forest for anomaly detection in cybersecurity. The process involves data collection, preprocessing, model training, evaluation, and comparison with other methods.

1. Data Collection The study utilizes a secondary dataset called UNSW-NB15, a benchmark dataset for network intrusion detection. The dataset contains labeled records of normal and malicious network traffic, capturing a variety of cyber threats such as DoS and Backdoor attacks.

2. Model Training A Random Forest classifier is trained on the preprocessed dataset. Key training parameters include:

- **Number of Trees:** 100 decision trees are used to balance performance and computational efficiency.
- **Criterion:** Gini impurity is employed to measure the quality of splits in decision trees.
- **Max Features:** The model considers the square root of the total number of features at each split to prevent overfitting.

3. Model Evaluation The trained model is evaluated using the following metrics:

- **Accuracy:** Measures the proportion of correctly classified instances.
- **Precision & Recall:** Assess how well the model differentiates between normal and malicious traffic.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

- **F1-Score:** Provides a balance between precision and recall for better anomaly detection assessment.
- **Confusion Matrix:** Visualizes the model's performance in identifying true positives, false positives, true negatives, and false negatives.

4. Comparison with Other Methods The performance of Random Forest is compared with an AutoEncoder-based anomaly detection model. While Random Forest provides better interpretability and real-time usability, the AutoEncoder model offers advantages in detecting novel anomalies without labeled data.

This methodology ensures a systematic approach to evaluating the effectiveness of Random Forest in cybersecurity anomaly detection while highlighting key comparisons with alternative model i.e. AutoEncoder.

IV. IMPLEMENTATION AND RESULTS

Model Training and Evaluation

A **Random Forest classifier** was trained with 100 decision trees. The following metrics were used to evaluate performance:

Classification Report:

Class	Precision	Recall	F1-Score	Support
Backdoor	0.98	0.08	0.14	1746
DoS	0.88	1	0.94	12264
Accuracy			0.88	14010
Macro Avg	0.93	0.54	0.54	14010
Weighted Avg	0.90	0.88	0.84	14010

Table 1: Classification Report of Random Forest Algorithm

Accuracy: 0.8845824411134904

Confusion Matrix Analysis: We present a confusion matrix to visualize prediction performance, highlighting misclassifications and areas for improvement.

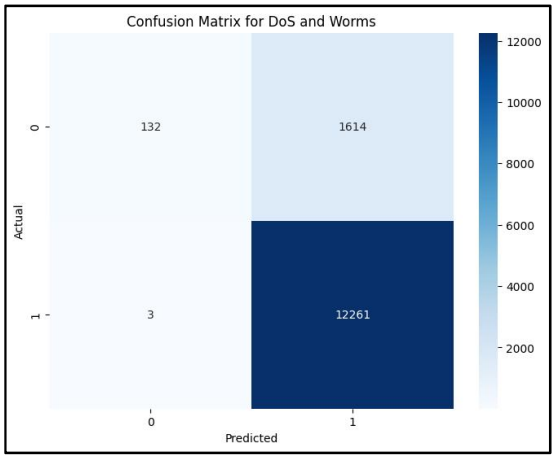

Fig 1: Confusion matrix for Random Forest classification algorithm

V. COMPARISON WITH OTHER METHODS

We compared RF with an AutoEncoder-based traffic anomaly detection model. While AutoEncoders can capture intricate data patterns, they require significantly more computational resources and training data. The AutoEncoder approach leverages unsupervised learning, where the model reconstructs normal traffic patterns and identifies anomalies based on reconstruction errors.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

Random Forest vs. AutoEncoder-Based Model

Metric	Random Forest	AutoEncoder
Accuracy	0.88	1
Precision	0.88	1
Recall	1	1
F1-Score	0.94	1
Training Time	30 seconds	10 minutes
Data Preprocessing	Minimal	StandardScaler
Dataset	UNSW-NB15	UNSW-NB15

Table 2: Random Forest vs. AutoEncoder-Based Model

AutoEncoder Model Accuracy graph :

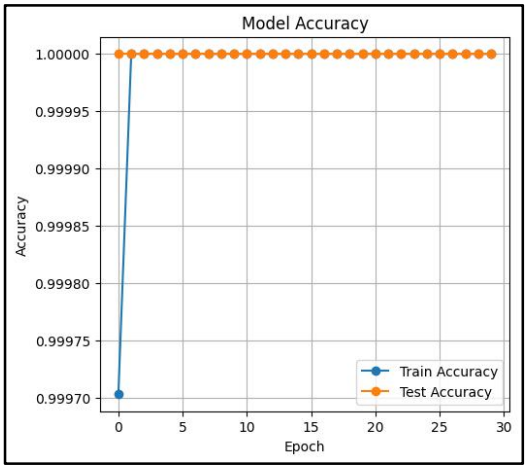

Fig 2: AutoEncoder Model Accuracy graph

AutoEncoder Model Loss graph :

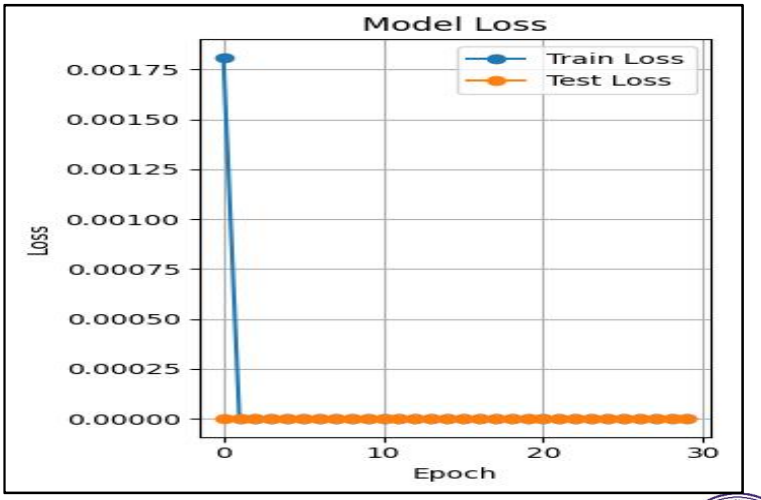

Fig 3: AutoEncoder Model Loss graph

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

The AutoEncoder-based model may provide advantages in detecting complex attack patterns, especially novel or unseen anomalies. However, it requires more computational resources, careful tuning of the anomaly detection threshold, and is generally less interpretable compared to RF. RF remains an optimal choice for real-time applications where speed and transparency are critical.

VI. CONCLUSION AND FUTURE DIRECTIONS

Our study demonstrates that **Random Forest is a powerful tool for anomaly detection** in cybersecurity. It offers high accuracy, interpretability, and efficiency. However, deep learning models provide slight improvements in accuracy at the cost of increased computational complexity.

Future Work

The AutoEncoder Model used in the study works on labeled data wherein the same can be explored using unlabeled data which might help to reduce the training time. Also, future research should explore hybrid models that combine Random Forest with deep learning techniques such as AutoEncoders to enhance anomaly detection capabilities. Additionally, improving feature selection techniques using advanced statistical methods or deep feature extraction could refine the model's accuracy and efficiency. Deploying RF-based models in real-time cybersecurity monitoring systems is another crucial area, ensuring that these models can handle large-scale network traffic dynamically. Further studies should also investigate AutoEncoder-based anomaly detection models to extract useful aspects for integration with RF. Finally, research should address computational efficiency concerns by optimizing RF for high-speed processing and low-latency environments, making it more viable for real-world cybersecurity applications.

REFERENCES

[1]. Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). *2015 Military Communications and Information Systems Conference (MilCIS)*, 1–6. https://doi.org/10.1109/MilCIS.2015.7348942

[2]. Moustafa, N., & Slay, J. (2016). The evaluation of network anomaly detection systems: Statistical analysis of the UNSW-NB15 dataset and the comparison with the KDD99 dataset. *Information Security Journal: A Global Perspective*, 25(1–3), 18–31. https://doi.org/10.1080/19393555.2015.1125974

[3]. Moustafa, N., & Slay, J. (2017). Novel geometric area analysis technique for anomaly detection using trapezoidal area estimation on large-scale networks. *IEEE Transactions on Big Data*, 5(4), 481–494. https://doi.org/10.1109/TBDATA.2017.2715165

[4]. Moustafa, N., & Slay, J. (2017). Big data analytics for intrusion detection system: Statistical decision-making using finite Dirichlet mixture models. In D. V. Kalashnikov, S. Mehrotra, & N. Venkatasubramanian (Eds.), *Data Analytics and Decision Support for Cybersecurity* (pp. 127–156). Springer. https://doi.org/10.1007/978-3-319-59439-2_7

[5]. Sarhan, M., Layeghy, S., Moustafa, N., & Portmann, M. (2020). NetFlow datasets for machine learning-based network intrusion detection systems. In S. Nepal, M. Pathan, & R. Buyya (Eds.), *Big Data Technologies and Applications: 10th EAI International Conference, BDTA 2020, and 13th EAI International Conference on Wireless Internet, WiCON 2020, Virtual Event, December 11, 2020, Proceedings* (p. 117). Springer. https://doi.org/10.1007/978-3-030-69990-4_9

[6]. Zhang, Y., Zhang, S., & Wei, X. (2019). A deep learning-based intrusion detection system for industrial control networks. *Journal of Network and Computer Applications*, 135, 33–40. https://doi.org/10.1016/j.jnca.2019.03.019

[7]. Wang, L., Li, J., & Zhao, S. (2020). A hybrid machine learning model for network anomaly detection. *International Journal of Computer Science and Network Security*, 20(2), 10–18.

[8]. Raza, M., & Abid, M. (2021). Comparative analysis of network intrusion detection techniques. *Journal of Computer Networks and Communications*, 2021, Article 6672824. https://doi.org/10.1155/2021/6672824

[9]. Wu, Y., Li, S., & Chen, X. (2022). Optimizing intrusion detection system using machine learning algorithms: A survey. *Computers, Materials & Continua*, 70(1), 1055–1070. https://doi.org/10.32604/cmc.2022.017989

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

[10]. Hu, X., Chen, T., & Liu, X. (2018). A novel feature selection method for anomaly-based network intrusion detection. *Journal of Computer Security*, 26(5), 621–637. https://doi.org/10.3233/JCS-171045

[11]. Kumar, S., Kumar, P., Tripathi, R., & Gupta, M. P. (2021). Research trends in network-based intrusion detection systems: A review. *IEEE Access*, 9, 157761–157786. https://doi.org/10.1109/ACCESS.2021.3129785

[12]. Songma, S., Netharn, W., & Lorpunmanee, S. (2024). Extending network intrusion detection with enhanced particle swarm optimization techniques. *arXiv preprint* arXiv:2408.07729. https://doi.org/10.48550/arXiv.2408.07729

[13]. Maseer, Z. K., Yusof, R., Al-Bander, B., Saif, A., & Kadhim, Q. K. (2023). Meta-analysis and systematic review for anomaly network intrusion detection systems: Detection methods, dataset, validation methodology, and challenges. *arXiv preprint* arXiv:2308.02805. https://doi.org/10.48550/arXiv.2308.02805

[14]. Talukder, M. A., Hasan, K. F., Islam, M. M., Uddin, M. A., Akhter, A., Yousuf, M. A., Alharbi, F., & Moni, M. A. (2022). A dependable hybrid machine learning model for network intrusion detection. *arXiv preprint* arXiv:2212.04546. https://doi.org/10.48550/arXiv.2212.04546

[15]. Tauscher, Z., Jiang, Y., Zhang, K., Wang, J., & Song, H. (2021). Learning to detect: A data-driven approach for network intrusion detection. *arXiv preprint* arXiv:2108.08394. https://doi.org/10.48550/arXiv.2108.08394