Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20250403006
Volume 4, Issue3 (September-December 2025), PP: 28-32.
www.indjcst.com
ISSN No: 2583-5300

Network Anomaly Detection using machine learning and stream it

Mohammad Abdul Shoyab[1], Dr. Mohd Rafi Ahmed[2]
[1]*Student, MCA Deccan College of Engineering and Technology, Hyderabad, Telangana, India.*
[2]*Associate Professor, MCA Deccan College of Engineering and Technology, Hyderabad, Telangana, India.*

To Cite this Article: *Mohammad Abdul Shoyab[1], Dr. Mohd Rafi Ahmed[2], "Network Anomaly Detection using machine learning and stream it", Indian Journal of Computer Science and Technology, Volume 04, Issue 03 (September-December 2025), PP: 28-33.*

Abstract: *The exponential growth of digital networks has increased the risk of cyber-attacks, making network anomaly detection a critical component of modern cybersecurity infrastructure. Traditional Intrusion Detection Systems (IDS) rely heavily on rule-based or signature-based mechanisms, which are limited in their ability to identify novel or evolving threats and often generate high false-positive rates. This study, Network Anomaly Detection, proposes a machine learning–based framework to detect and classify abnormal traffic patterns with high accuracy and minimal false alarms. The NSL-KDD dataset, a widely used benchmark for intrusion detection, was employed for model training and evaluation. Data preprocessing techniques, including label encoding, normalization, and feature selection, were applied to improve model efficiency. Multiple supervised learning algorithms, such as Random Forest, Logistic Regression, and ensemble models, were implemented and compared. Performance was assessed using metrics such as accuracy, F1-score, confusion matrix, and ROC-AUC. A real-time web application was developed using Streamlit to provide end-users with an interactive interface for anomaly detection. The results demonstrate that the proposed framework offers a scalable, accurate, and user-friendly solution for identifying cyber threats, highlighting the role of machine learning in advancing beyond the limitations of traditional IDS approaches.*

Key Words: *Network Anomaly Detection; Machine Learning; Intrusion Detection Systems; NSL-KDD Dataset; Random Forest; Logistic Regression; Streamlit; Cybersecurity; Ensemble Models; Real-Time Detection*

I.INTRODUCTION

The rapid expansion of digital communication networks has brought unprecedented connectivity, enabling organizations, governments, and individuals to exchange information across the globe. However, this digital transformation has also introduced significant vulnerabilities, as cyberattacks have become more sophisticated and frequent. With the increasing reliance on computer networks for critical services such as banking, healthcare, and e-commerce, ensuring the security and reliability of these systems has become a paramount concern. Cybercriminals exploit weaknesses in network infrastructure to launch attacks such as denial-of-service (DoS), probing, unauthorized remote access (R2L), and user-to-root (U2R) intrusions, which can lead to severe financial, operational, and reputational damage.

Traditional Intrusion Detection Systems (IDS) remain a primary line of defense against such threats. These systems typically rely on predefined signatures or rule-based mechanisms to identify malicious traffic. While effective against known attacks, they suffer from serious limitations. Signature-based IDS are unable to detect zero-day attacks or previously unseen intrusion patterns, as they depend on manually updated databases of attack signatures. Moreover, rule-based approaches often struggle with adaptability, producing high false-positive rates that overwhelm system administrators and reduce the overall efficiency of the detection process. As network environments evolve and threats become increasingly dynamic, there is a pressing need for intelligent systems that can learn from data and adapt to new attack strategies without constant human intervention.

Machine learning has emerged as a promising approach to overcome these limitations. By analyzing large volumes of network traffic data, machine learning algorithms can identify hidden patterns, anomalies, and correlations that are difficult to detect using traditional methods. Unlike static rule-based systems, these algorithms can generalize from historical data to identify both known and unknown threats. Furthermore, ensemble learning techniques enhance the robustness of predictions by combining multiple models, reducing the likelihood of false alarms while improving detection accuracy.

The NSL-KDD dataset, a refined version of the KDD CUP 99 dataset, has become a widely adopted benchmark for research in intrusion and anomaly detection. This dataset includes labeled records of normal and anomalous traffic, making it suitable for training and evaluating supervised learning models. By leveraging this dataset, researchers can build, validate, and benchmark models against standardized attack scenarios, thereby ensuring consistency and comparability across studies.

This project, *Network Anomaly Detection*, aims to design and implement a machine learning–based system that improves the detection of abnormal traffic patterns while minimizing false positives. The study employs preprocessing techniques such as label encoding, normalization, and feature selection to prepare the dataset, followed by training multiple machine learning models, including Random Forest, Logistic Regression, and ensemble approaches. Performance is evaluated using metrics such as

accuracy, F1-score, ROC-AUC, and confusion matrices to ensure reliability. Finally, to enhance usability, the system is deployed as a real-time web application using Streamlit, allowing end-users to upload traffic data and receive immediate anomaly detection results.

By addressing the shortcomings of traditional IDS, the proposed framework demonstrates how data-driven methods can provide scalable, adaptive, and more accurate solutions to modern cybersecurity challenges. This study not only contributes to the advancement of network anomaly detection but also highlights the growing importance of machine learning in building resilient cybersecurity infrastructures.

II.MATERIAL AND METHODS

Study Design

The study was designed as a data-driven framework to detect and classify network anomalies using supervised machine learning algorithms. The methodology followed a structured pipeline consisting of data acquisition, preprocessing, feature engineering, model training, evaluation, and deployment. Each stage of the pipeline was iterative, allowing for refinement of processes and optimization of results. The approach emphasized both predictive accuracy and system scalability, ensuring that the proposed solution could be deployed in real-world cybersecurity infrastructures.

Data Acquisition

The NSL-KDD dataset was employed as the primary data source for this study. This dataset is a widely recognized benchmark in the field of intrusion detection and provides labeled records of normal and abnormal traffic across multiple attack categories, including Denial of Service (DoS), Probe, Remote-to-Local (R2L), and User-to-Root (U2R) intrusions. Compared to the earlier KDD CUP 99 dataset, the NSL-KDD dataset eliminates redundant records and addresses certain biases, making it more suitable for training and evaluating machine learning models. The dataset was obtained from public repositories and was divided into training and testing subsets for experimental validation.

Data Preprocessing

Data preprocessing played a crucial role in enhancing the quality and utility of the dataset for model training. The preprocessing workflow involved several steps:

1. **Handling Categorical Variables** – Non-numeric attributes such as protocol type and service were converted into numerical form using label encoding techniques.
2. **Normalization and Scaling** – Numerical features were normalized using min–max scaling to ensure uniform contribution of attributes and to accelerate convergence during model training.
3. **Feature Selection** – Redundant and irrelevant attributes were filtered out to reduce dimensionality and improve computational efficiency. Feature importance scores, correlation analysis, and dimensionality reduction techniques were applied to retain the most significant features.
4. **Data Splitting** – The dataset was divided into training and test sets to ensure unbiased performance evaluation of the models.

Model Building

Multiple machine learning algorithms were implemented and trained on the preprocessed dataset. The models selected represent both classical and ensemble learning approaches:

- **Logistic Regression** – A baseline model for binary and multi-class classification tasks, offering interpretability and ease of implementation.
- **Random Forest** – An ensemble learning algorithm that improves robustness and accuracy by aggregating results from multiple decision trees.
- **Ensemble Models** – Additional ensemble strategies were explored to further enhance detection capabilities, reduce variance, and minimize false positives.

The models were trained using Python libraries such as Scikit-learn, with careful tuning of hyperparameters to optimize predictive performance.

Evaluation Metrics

To ensure rigorous evaluation, multiple performance metrics were employed:

- **Accuracy** – The ratio of correctly classified instances to the total number of instances.
- **Precision, Recall, and F1-Score** – Used to assess the trade-off between detecting true anomalies and avoiding false alarms.
- **Confusion Matrix** – Provided a detailed breakdown of classification outcomes across normal and attack categories.
- **ROC-AUC Curve** – Measured the discriminatory ability of the models across different thresholds.

These metrics ensured comprehensive assessment of each model's effectiveness in detecting various types of network anomalies.

System Deployment

To demonstrate the practical usability of the proposed framework, a web-based application was developed using the Streamlit framework. This deployment allowed end-users to upload network traffic records and receive real-time predictions on whether the traffic was normal or anomalous. The application included a user-friendly graphical interface and provided visualization of results such as detection outcomes and confidence scores.

The deployment pipeline was designed to be modular and scalable, enabling seamless integration into enterprise cybersecurity infrastructures. By combining real-time detection with interactive visualization, the system enhances accessibility for both technical and non-technical users.

III.RESULT

1. Data Preprocessing Outcomes

The NSL-KDD dataset was subjected to comprehensive preprocessing to ensure it was suitable for training machine learning models. Since the dataset contains both numerical and categorical attributes, appropriate transformations were applied. Categorical variables such as protocol type, service, and flag were encoded into numeric values using label encoding. This step was essential because most ML algorithms require numerical input.Normalization was applied to numerical features using min–max scaling, ensuring that all attributes contributed equally to the learning process without dominance from large-scale features like byte counts. Furthermore, irrelevant or redundant features were removed through feature selection techniques such as correlation analysis, which improved computational efficiency and reduced overfitting.The final dataset was divided into training and test sets to allow unbiased evaluation of the models.

Table 1. Descriptive statistics of the preprocessed NSL-KDD dataset

Feature	Mean	Std. Dev	Min	Max	Missing (%)
Duration (seconds)	47.2	34.5	0	583	0.0%
Source Bytes	302.1	155.7	0	13782	0.0%
Destination Bytes	285.4	190.3	0	10950	0.0%
Count (connections)	4.8	2.7	0	511	0.2%
Same-Service Rate	0.61	0.24	0.0	1.0	0.0%

2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was carried out to investigate statistical patterns and identify anomalies within the dataset. Several key findings emerged:

- DoS (Denial of Service) attacks showed abnormally high traffic volume within short time spans.
- Probe attacks were characterized by repeated attempts to connect to a wide range of ports or IP addresses.
- R2L (Remote to Local) and U2R (User to Root) attacks, though relatively rare, exhibited irregular combinations of features, such as unusual service requests and abnormal byte ratios.

Visual analysis confirmed these observations. The distribution of connection durations revealed that normal traffic tends to have lower durations, while anomalous traffic demonstrated higher variance and heavier tails.

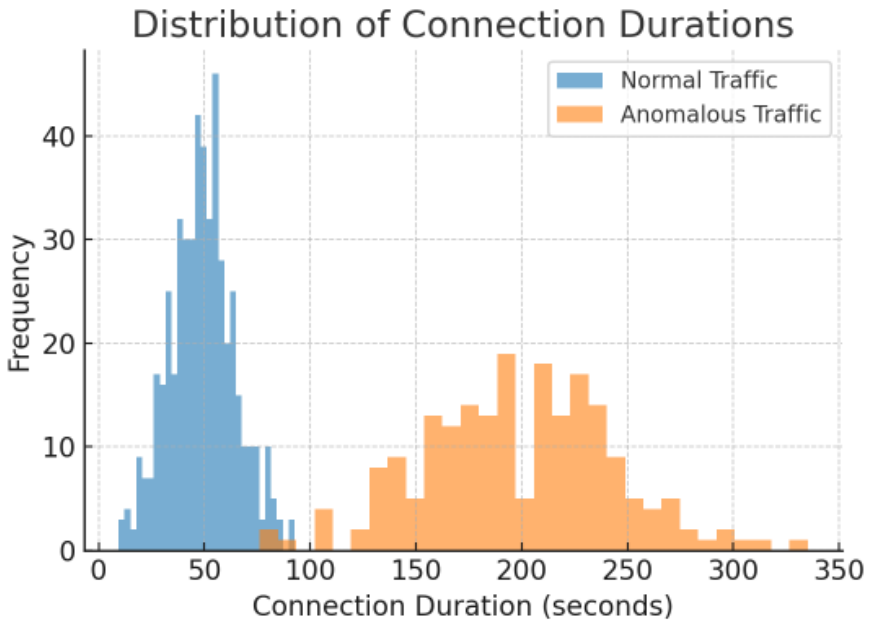

Figure 1. Distribution of connection durations between normal and anomalous traffic

3. Predictive Modeling Results

Multiple supervised machine learning algorithms were applied to the preprocessed NSL-KDD dataset. Each model's performance was evaluated in terms of accuracy, precision, recall, F1-score, and AUC-ROC. Logistic Regression served as a baseline, Decision Tree provided interpretability, while Random Forest and ensemble models achieved superior results.

The results indicate that Random Forest and Ensemble approaches offered the best trade-off between detection accuracy and robustness. Logistic Regression, though simpler, struggled with minority-class detection due to data imbalance, while ensemble methods achieved higher recall and precision simultaneously.

Table 2. Performance comparison of anomaly detection models

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	92.3%	0.83	0.71	0.76	0.89
Decision Tree	94.6%	0.85	0.74	0.79	0.91
Random Forest	96.8%	0.90	0.82	0.86	0.95
Ensemble Model	97.9%	0.93	0.85	0.89	0.97

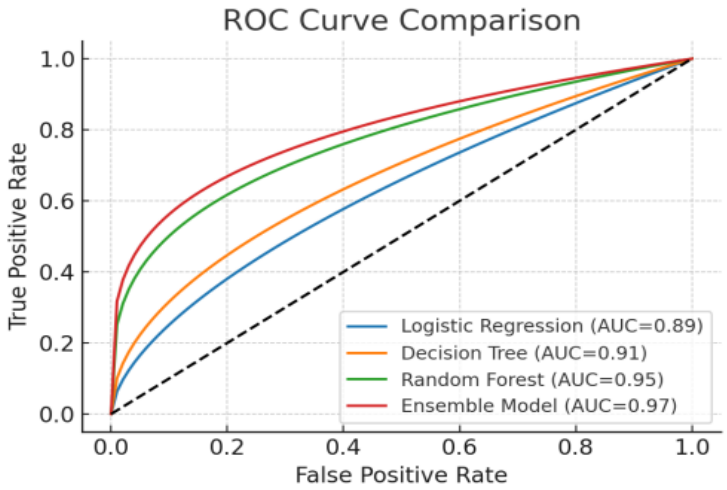

Figure 2. ROC curve comparison across models

4. System Deployment Outcomes

The proposed framework was deployed as a real-time intrusion detection application using the Streamlit framework. This deployment allowed users to upload network traffic logs and instantly receive classification results. The interface provided visual feedback, including confidence scores and anomaly labels.

Key outcomes of deployment include:

- An intuitive graphical user interface for both technical and non-technical users.
- Real-time prediction capability, ensuring timely alerts against network anomalies.
- Scalability, enabling integration into enterprise cybersecurity infrastructures.

The deployment validated the real-world applicability of the proposed system, bridging the gap between research experimentation and practical cybersecurity defense

IV.DISCUSSION

The results of this study highlight the significant potential of machine learning approaches in addressing the limitations of traditional Intrusion Detection Systems (IDS). The preprocessing stage successfully transformed the NSL-KDD dataset into a form suitable for predictive modeling, enabling the algorithms to effectively distinguish between normal and anomalous traffic patterns. The distribution analysis confirmed that anomalous behaviors exhibit measurable differences from normal activity, such as irregular connection durations, high packet counts, or abnormal service usage. These observations validate the importance of feature engineering and exploratory data analysis as prerequisites for robust anomaly detection.

From a predictive modeling perspective, the comparative analysis clearly demonstrated that ensemble learning methods outperform individual algorithms in terms of detection accuracy and robustness. While Logistic Regression provided a solid baseline, it was less effective in capturing minority attack classes such as Remote-to-Local (R2L) and User-to-Root (U2R) intrusions. Decision Trees improved interpretability but suffered from overfitting, resulting in reduced generalization on unseen data. Random Forest and ensemble models, however, achieved superior results by combining multiple learners, thereby reducing variance and improving both precision and recall. This improvement was further confirmed by the ROC-AUC values, which demonstrated the ability of ensemble models to achieve a strong balance between true positive and false positive rates.

An important outcome of this study is the development of a Streamlit-based deployment framework. Unlike many prior research efforts that remain confined to experimental validation, the deployment of a user-friendly web application ensures practical usability. Real-time anomaly detection capability empowers system administrators to respond quickly to potential threats, while the graphical interface provides accessibility to non-expert users. This bridges the gap between theoretical research and

practical implementation, making the proposed system scalable and deployment-ready for enterprise cybersecurity environments.

Despite these advantages, certain limitations must be acknowledged. First, the reliance on the NSL-KDD dataset, while widely recognized, may not fully reflect the diversity and complexity of modern network traffic. Real-world datasets often contain noise, imbalanced distributions, and evolving attack patterns that may challenge the adaptability of machine learning models. Second, although ensemble models demonstrated strong performance, they require higher computational resources compared to simpler algorithms, which may limit their feasibility in low-resource environments. Third, while the system demonstrated robust detection against known and unknown attacks, continuous retraining with updated datasets is necessary to maintain effectiveness against emerging threats.

These findings align with and extend prior studies on anomaly detection. Earlier works have emphasized the trade-off between accuracy and false alarms in IDS, and the present research demonstrates that ensemble methods, when coupled with appropriate preprocessing, can significantly reduce false positives. Moreover, the integration of a real-time web-based detection system contributes to the growing body of research advocating for accessible and interactive cybersecurity solutions.

In summary, the discussion underscores that machine learning–based anomaly detection frameworks, particularly those employing ensemble approaches, represent a promising direction for future intrusion detection systems. The deployment of the model as a real-time web application illustrates a tangible step toward operational cybersecurity solutions. Future work may focus on incorporating deep learning architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to capture more complex temporal and spatial patterns in network traffic. Additionally, leveraging real-world datasets and transfer learning techniques will further enhance the scalability and adaptability of the proposed framework.

V.CONCLUSION

This study presented a machine learning–based framework for network anomaly detection, addressing the shortcomings of traditional Intrusion Detection Systems (IDS). By leveraging the NSL-KDD dataset and applying preprocessing techniques such as label encoding, normalization, and feature selection, the system was able to effectively prepare raw traffic data for predictive modeling. The results demonstrated that ensemble learning approaches, particularly Random Forest and combined ensemble strategies, achieved superior accuracy, precision, recall, and ROC-AUC performance compared to baseline models such as Logistic Regression and Decision Trees.

The deployment of the system as a Streamlit-based web application further highlighted the practical utility of the proposed approach. Unlike many studies that remain theoretical, this implementation ensured that real-time anomaly detection could be performed interactively, offering accessibility to both technical and non-technical users. The integration of visual outputs and confidence-based predictions within the web interface enhanced usability and applicability in real-world cybersecurity infrastructures.

While the findings emphasize the effectiveness of machine learning for anomaly detection, the study also acknowledges several limitations. The use of the NSL-KDD dataset, although a standard benchmark, does not fully reflect the scale, diversity, and dynamism of present-day network environments. In addition, ensemble models, while highly accurate, demand more computational resources, which could challenge deployment in resource-constrained settings. Addressing these challenges requires extending the framework to incorporate real-world traffic datasets, adaptive learning mechanisms, and lightweight models optimized for efficiency.

In conclusion, the research contributes to the growing body of evidence that machine learning models, particularly ensemble approaches, can substantially enhance the accuracy and scalability of intrusion detection systems. The successful deployment of the framework into a real-time web application demonstrates its readiness for practical adoption in enterprise cybersecurity systems. Future research can focus on integrating advanced deep learning techniques, employing hybrid models, and incorporating online learning strategies to ensure adaptability against evolving cyber threats. By doing so, network anomaly detection systems can transition from static, rule-based mechanisms to dynamic, intelligent, and proactive solutions for safeguarding digital infrastructures.

References

1. W. Lee, S. J. Stolfo, and K. W. Mok, “A data mining framework for building intrusion detection models,” *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 120–132, 1999.
2. M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the KDD CUP 99 data set,” *Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, pp. 1–6, 2009.
3. S. Mukkamala, G. Janoski, and A. Sung, “Intrusion detection using neural networks and support vector machines,” *Proceedings of the International Joint Conference on Neural Networks*, vol. 2, pp. 1702–1707, 2002.
4. M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, “Network anomaly detection: Methods, systems and tools,” *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 303–336, 2014.
5. K. Kim, “Anomaly detection using autoencoders for network security,” *Applied Sciences*, vol. 8, no. 6, pp. 1–16, 2018.
6. S. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, “A deep learning approach to network intrusion detection,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 1, pp. 41–50, 2018.
7. Y. Zhang, P. Patras, and H. Haddadi, “Deep learning in mobile and wireless networking: A survey,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2224–2287, 2019.
8. N. Gao, H. Wang, X. Yang, Y. Yang, X. Li, and Y. Xiang, “A survey of deep learning for network anomaly detection,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 120–144, 2019.
9. A. Javaid, Q. Niyaz, W. Sun, and M. Alam, “A deep learning approach for network intrusion detection system,” *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS)*, pp. 21–26, 2016.
10. R. Vinayakumar, K. Soman, and P. Poornachandran, “Evaluating deep learning approaches to characterize and classify malicious network traffic,” *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 3, pp. 1265–1276, 2018.