iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 Issue: I Month of publication: January 2026

DOI: https://doi.org/10.22214/ijraset.2026.77221

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue I Jan 2026- Available at www.ijraset.com

Explainable Network Intrusion Detection Using Random Forest and SHAP on the CICIDS2017 Dataset

Ashwitha C Shetty
Department of Computer Science Engineering, (Internet of Things& Cyber Security Block Chain Technology), K S Institute of Technology, Bengaluru , India

Abstract: Network intrusion detection plays a vital role in protecting modern computer networks and IoT environments from cyber threats. Traditional machine learning approaches often focus on improving detection accuracy without providing insight into decision-making processes. This paper proposes an explainable intrusion detection framework using a Random Forest classifier trained on the CICIDS2017 dataset. The dataset includes realistic benign and malicious network traffic, enabling effective model evaluation. The proposed model achieves an accuracy of 92% in distinguishing benign and attack traffic. To enhance transparency and interpretability, SHAP (SHapley Additive exPlanations) is employed to analyze feature contributions influencing predictions. Experimental results demonstrate that flow-based and packet-level features significantly impact attack detection. The integration of explainable AI improves trust and usability of machine learning-based intrusion detection systems in real-world cyber security applications.

Keywords: Intrusion Detection, Cyber Security, Machine Learning, Random Forest, Explainable AI, SHAP, CICIDS2017

I. INTRODUCTION

With the rapid growth of internet-connected systems and IoT devices, cyber attacks have become increasingly sophisticated and frequent. Network intrusion detection systems (IDS) are designed to monitor network traffic and identify malicious activities. Traditional IDS techniques rely on signature-based detection, which fails to detect unknown or evolving attacks. Machine learning-based IDS offers improved detection capabilities by learning patterns from network traffic data. However, many machine learning models operate as black boxes, making it difficult to understand how decisions are made. This lack of interpretability reduces trust and limits adoption in critical cyber security environments. To address this issue, this work proposes an explainable intrusion detection approach using a Random Forest classifier combined with SHAP-based feature explanation. The CICIDS2017 dataset is used to evaluate the effectiveness of the proposed framework.

II. LITERATURE REVIEW

Several studies have explored machine learning techniques such as Support Vector Machines, Decision Trees, Random Forests, and Deep Learning for intrusion detection . Random Forest models have demonstrated strong performance due to their robustness and ability to handle high-dimensional data. Recent research has highlighted the importance of explainable AI in cyber security to understand and validate model decisions. SHAP has emerged as a popular method for interpreting complex machine learning models by assigning feature importance values based on cooperative game theory. Despite this, limited studies combine intrusion detection with explain ability using realistic datasets such as CICIDS2017. Prior studies using CICIDS2017 have applied SVM, Deep Neural Networks, and ensemble methods for intrusion detection. However, many of these works focus solely on accuracy and lack interpretability analysis. Recent explainable AI approaches highlight the need for feature-level transparency, which remains underexplored in intrusion detection contexts.

III. DATASET DESCRIPTION

The experiments in this study utilize the CICIDS2017 dataset, developed by the Canadian Institute for Cyber security (CIC). The dataset is designed to closely resemble real-world network traffic by capturing both benign and malicious activities in a controlled environment. Unlike older intrusion detection datasets, CICIDS2017 includes modern attack scenarios and realistic traffic patterns, making it suitable for evaluating machine learning-based intrusion detection systems. Each network flow is represented using 78 extracted statistical features describing packet size, timing, and traffic behaviour.

A. Dataset Generation Environment

The CICIDS2017 dataset was generated in a test bed that simulated real organizational network behaviour. The environment consisted of multiple machines configured as attackers, victims, and normal users. Network traffic was captured using packet sniffing tools, and flow-based features were extracted using CIC Flow Meter. This approach ensures that the dataset contains both low-level packet information and high-level flow characteristics. The dataset was collected over a period of five days, with each day representing different types of network behaviour:

1) Monday: Normal (benign) traffic
2) Tuesday to Friday: Various attack scenarios mixed with normal traffic

IV. METHODOLOGY

A. Data Pre-processing

1) Removal of irrelevant metadata
2) Label encoding (Benign = 0, Attack = 1)
3) Handling missing and infinite values
4) Feature scaling using Standard Scaler
5) Stratified train-test split (80% training, 20% testing)

A Random Forest classifier was selected due to its robustness and ability to handle large-scale network data. The model was trained using 50 decision trees with controlled depth to balance performance and computational efficiency. Additionally a Random Forest classifier was selected and trained as the primary detection model due to its robustness and performance on high-dimensional network traffic data on the dataset.

Table I — Random Forest Parameters

Parameter	Value
Trees	50
Max Depth	15
Train/Test Split	80/20

.

V. RESULTS AND DISCUSSION

This section presents the experimental results obtained from the proposed Random Forest–based intrusion detection framework and discusses the observed performance. The evaluation focuses on classification effectiveness as well as the interpretability of model decisions.

A. Experimental Setup

The experiments were conducted using the CICIDS2017 dataset, consisting of benign traffic from *Benign-Monday* and DDoS attack traffic from *DDoS-Friday*. After pre processing and feature scaling, the dataset was split into training and testing sets using an 80:20 stratified split to preserve class distribution. A Random Forest classifier with 50 trees and controlled depth was employed to balance detection performance and computational efficiency.

B. Classification Performance

The trained model was evaluated using standard performance metrics, including accuracy, precision, recall, F1-score, and confusion matrix. These metrics provide a comprehensive understanding of the model's ability to distinguish between benign and malicious traffic.

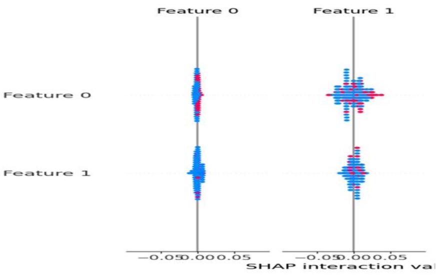

Fig 1. SHAP summary plot illustrating the impact of feature values on intrusion detection predictions

Table II —Classification Performance Metrics.

Metric	***Value***
Accuracy	92%
Precision	93%
Recall	81%
F1-score	87%

VI. CONCLUSION AND FUTURE WORK

This study presented an explainable intrusion detection framework using a Random Forest classifier trained on the CICIDS2017 dataset. The model achieved 92% accuracy while effectively distinguishing malicious from benign traffic. SHAP-based explanations enhanced transparency by revealing feature contributions influencing predictions. Results confirm that meaningful network flow features drive model decisions, making the approach suitable for practical cyber security applications. Overall, the results validate the effectiveness and interpretability of the proposed intrusion detection approach, making it suitable for practical cyber security and IoT network monitoring applications.

A. Future Improvements may Include

Future work will focus on extending the proposed system to detect multiple attack categories, performing real-time deployment testing, integrating deep learning approaches, evaluating performance on newer datasets, and optimizing the model for deployment on IoT edge devices.

REFERENCES

[1] Canadian Institute for Cybersecurity, “CICIDS2017 Dataset,” University of New Brunswick, 2017.
[2] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[3] S. M. Lundberg and S. I. Lee, “A Unified Approach to Interpreting Model Predictions,” Advances in Neural Information Processing Systems (NeurIPS), 2017.
[4] R. Sommer and V. Paxson, “Outside the Closed World: On Using Machine Learning for Network Intrusion Detection,” IEEE Symposium on Security and Privacy, 2010.
[5] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization,” ICISSP, 2018
.