IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

AI-Based Cybersecurity Threat Detection System for Organizational Networks

Mr. Dinesh P, Perarasu P, Ram Varma S,Vigneshwaran S, Vijay J
Assistant Professor, Department of Computer Science and Engineering
Students, Department of Computer Science and Engineering
Anjalai Ammal Mahalingam Engineering College Kovilvenni, Tamil Nadu, India

Abstract: *Enterprise networks today generate a large amount of constantly evolving traffic, thus making it increasingly more difficult to manually monitor and investigate security events. Traditional intrusion detection systems struggle to detect some of the more complex or never-before-seen attack patterns. This paper presents the intelligent threat detection architecture, which consists of two modules: a detection module (which uses Machine Learning algorithms) and an explanation module (which utilises a Large Language Model (LLM)). The detection module uses Random Forest algorithms to classify network traffic based on features derived from the CIC-IDS2017 dataset. Once suspicious activity has been detected using this classification, it generates contextual explanations that provide analysts with an understanding of how the attack was perpetrated and its severity. By integrating automated detection with interpretability, this architecture enables analysts to more effectively interpret and understand the nature of the threats to their networks. Experimental studies show that while the Random Forest model achieves high classification performance, the anomaly-based model provides assistance in identifying unknown attack behaviour. The evidence from the LLM explanation will enhance communication with incident response teams, enabling them to make informed decisions. The demonstrated feasibility of the combination of Machine Learning with Language Models to develop interpretable and effective cybersecurity monitoring systems.*

Keywords: Intrusion Detection System, Network Security, Random Forest, Large Language Model.

I. INTRODUCTION

The monitoring of security-related activities by human operators has become increasingly challenging due to the volume of data transmitted across the networks associated with contemporary organisations. Typically, Intrusion Detection Systems (IDS) are employed to identify malicious behaviours, such as attacks, and abnormal traffic patterns within a given network environment. For instance, many traditional IDS utilize rule-based methodologies to identify malicious activities; however, these systems often encounter difficulties in detecting sophisticated attacks or those originating from previously unidentified actors.

Recent research findings suggest that using machine learning techniques can significantly improve the performance of IDS by leveraging historical network traffic data to learn patterns associated with legitimate and illegitimate user behaviours. Examples of machine learning methods that have been used with IDS include ensemble methods, such as the Random Forest algorithm, which are effective for handling very high-dimensional datasets and classifying items with high accuracy.

Advancements in IDS have also occurred recently, in part due to the availability of realistic datasets created specifically to support intrusion detection research. The CIC-IDS2017 dataset, developed by researchers at the Canadian Institute for Cybersecurity, contains labelled network traffic for various types of attacks; researchers can use this dataset to evaluate the accuracy of their IDS models. While machine learning methods can be successful, many current systems for detecting intrusion cannot explain their decisions, making it difficult for security analysts to analyse a model and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

respond appropriately. Recent advances in transformer-based language models have created a set of powerful tools for generating contextualised, human-readable explanations of how a structured piece of information, such as that produced by a machine learning model, was assessed [8]. This paper describes a framework for AI-based threat detection that consists of a Detection Module based on Machine Learning and an Explanation Module based on a large language model (LLM), which work together to classify and detect attacks against an organisation and provide human-readable contextual explanations of the detected attacks.

II. LITERATURE REVIEW

The field of security has been widely researched on the topic of intrusion detection systems. The initial methods included an approach that used signature-based intrusion detection systems. This type of system compares traffic patterns to known attacks. Signature-based methods are capable of detecting previously identified attack patterns but cannot detect new attacks or modified versions of those attacks.

As a result, researchers have begun to develop new methods to enhance the performance of intrusion detection systems using machine learning algorithms. Buczak and Guven have provided a thorough review of the data mining and machine-learning approaches being utilised for intrusion detection systems [4]. Based on their research, it is evident that classification of algorithms can be highly effective in improving the performance of intrusion detection systems when trained using data from representative datasets of network traffic.

Numerous researchers have investigated the use of ensemble learning approaches to improve the classification of data by combining multiple decision trees in a single classifier, such as the Random Forest algorithm developed by Breiman[2]. Another significant method for intrusion detection is through the application of Anomaly Detection. In particular, the Isolation Forest algorithm achieves an efficient method for finding anomalies (or Outliers) by recursively partitioning the feature space to isolate anomalous observations [3]. Standard references have also recently indicated that the Explainability of AI in Cybersecurity Systems should receive additional consideration. Explainable Artificial Intelligence (XAI) techniques should assist investigators in modelling the behaviour of models to help interpret and trust automated anomaly detection mechanisms [7]. Additionally, transformer architectures have significantly improved natural language processing (NLP) tasks. The transformer architecture, as proposed in the “Attention Is All You Need” [8] paper, provides an appropriate approach for utilising large language models to identify relationships between various entities in text. Therefore, it would be appropriate to combine security analytics with these large language models to assist in generating descriptive explanations following the detection of threats.

III. RESEARCH GAP

Although existing research has demonstrated the effectiveness of machine learning techniques in intrusion detection systems, several challenges still remain. Many traditional IDS solutions focus primarily on attack detection accuracy, while providing limited support for interpretability and contextual understanding of detected threats.

Most machine learning-based IDS models operate as black-box systems, producing predictions without clearly explaining the reasoning behind them. This lack of interpretability makes it difficult for security analysts to evaluate the credibility of the detected threats and respond appropriately.

Additionally, many systems rely on either supervised classification models or unsupervised anomaly detection models but rarely combine both approaches within a unified architecture. Furthermore, the integration of natural language generation techniques for explaining detected attacks remains relatively unexplored.

Therefore, there is a need for a framework that:

- Combines multiple machine learning techniques for robust detection.
- Provides human-readable explanations for detected attacks.
- Improves the interpretability of intrusion detection systems for cybersecurity analysts.

This research addresses these gaps by integrating machine learning-based threat detection with a Large Language Model (LLM)-based explanation module.

DOI: 10.48175/IJARSCT-33250

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

IV. PROPOSED FRAMEWORK

The proposed framework integrates machine learning-based threat detection with a language model-based explanation system to enhance the interpretability of intrusion detection results. The framework consists of two primary components:

A. Threat Detection Module

This module analyses network traffic using machine learning algorithms to classify traffic as normal or malicious.

B. Explanation Module

Once an attack is detected, the system uses a large language model to generate human-readable explanations describing:

- The detected attack type
- Key features contributing to the detection
- Possible security impacts
- Suggested mitigation actions

This integrated approach improves both detection performance and analyst understanding.

V. SYSTEM ARCHITECTURE

The AI-based Threat Detection Architecture is made up of several components that work together to analyse the data generated by network traffic to produce intelligence about potential threats. The architecture integrates a preprocessing pipeline, dual detection models, and a language model-based explanation interface, forming an end-to-end system for intelligent network security monitoring.

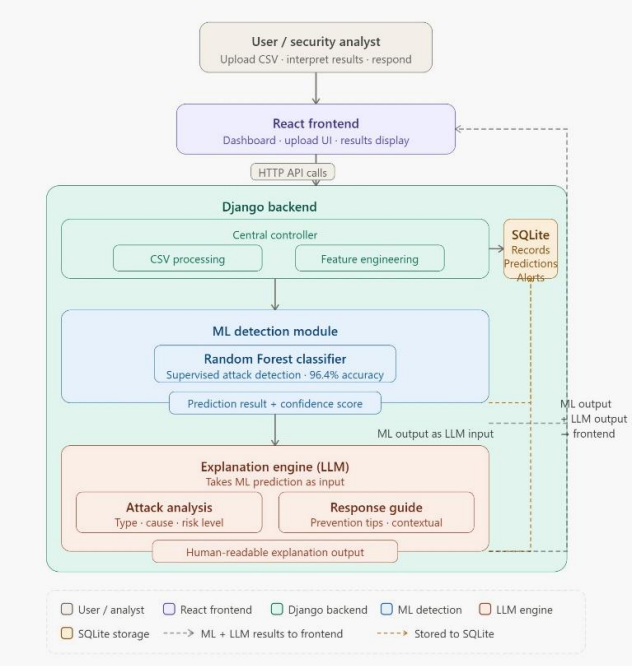

Fig. 1 System Architecture of the AI-Based Threat Detection Framework

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

VI. IMPLEMENTATION

The proposed system was implemented using a machine learning pipeline for traffic analysis and a language model interface for explanation generation.

A. Data Processing

- Dataset loading
- Data cleaning
- Feature normalization

B. Model Training

- Random Forest classifier for supervised attack detection

C. Prediction System

Incoming network traffic features are processed and passed to the trained detection models.

D. Explanation Generation

Once a threat is detected, the prediction output and related features are provided to the language model, which generates a detailed explanation of the attack. The system can be integrated into organisational network monitoring platforms for automated threat analysis.

VII. EXPERIMENT SETUP

This study used the CIC-IDS2017 dataset, which contains realistic network traffic including both benign and malicious activities.

A. Attack Types Included

- Denial of Service (DoS)
- Distributed Denial of Service (DDoS)
- Brute Force attacks
- Port Scanning attacks

B. Data Preprocessing Steps

- Removing incomplete records
- Normalising numerical values
- Converting categorical labels
- Feature selection

C. Training Process

The dataset was divided into a Training Set — used to train machine learning models — and a Testing Set — used to evaluate model performance. Evaluation metrics included Accuracy, Precision, and Recall.

VIII. RESULTS AND DISCUSSION

TABLE I: MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall
Random Forest	96.4%	95.9%	96.1%

The Random Forest classifier demonstrated superior performance in detecting known attacks due to its ensemble learning capability.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

The integration of the LLM-based explanation module further enhances system usability by providing clear descriptions of detected threats, enabling security analysts to understand attack scenarios more effectively.

The system was deployed and evaluated through a live web interface named Sentinel AI. Fig. 2 illustrates the Projects page, which serves as the entry point for organising network traffic scans. Each project displays key metrics including the total number of records processed, the number of detected threats, and the timestamp of the most recent scan. In the test project shown, 1,995 records were processed and 1,196 threats were identified within a 15-minute scan window, demonstrating the system's ability to handle large volumes of network traffic in near real-time.

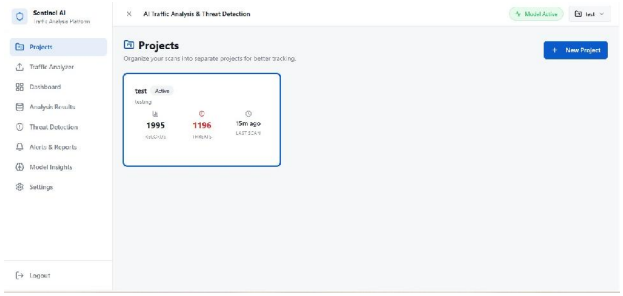

Fig. 2 Sentinel AI projects dashboard showing scan metrics

Fig. 3 shows the Traffic Analyser interface, which allows users to upload CIC-IDS2017-formatted CSV files containing 69 network traffic features. The drag-and-drop upload mechanism simplifies the process of submitting traffic data for analysis. Once a file is uploaded, the user initiates analysis by clicking the Analyse Traffic button, which triggers the Django backend to pass the data through the feature engineering pipeline and subsequently to the Random Forest classifier.

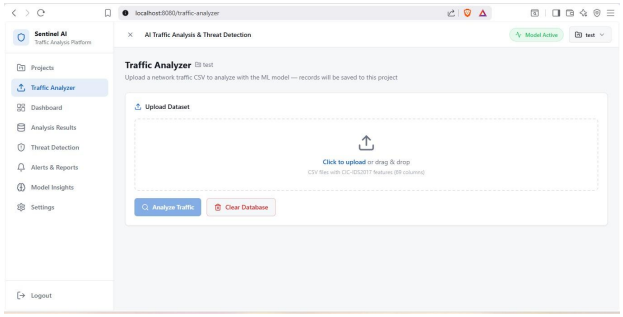

Fig. 3 Traffic Analyser interface for uploading CIC-IDS2017 CSV datasets

Fig. 4 presents the Analysis Overview dashboard, which provides a high-level summary of the ML model's output. For the test project, 2,000 records were processed, of which 1,200 were classified as attacks across 14 distinct attack types. The dashboard includes an Attack Type Breakdown bar chart and a Severity Distribution donut chart. The severity analysis revealed that 1,126 threats were classified as High severity, 28 as Medium, and 841 as Low, underscoring the prevalence of high-impact attacks such as DoS Hulk, PortScan, and DDoS in the evaluated dataset.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

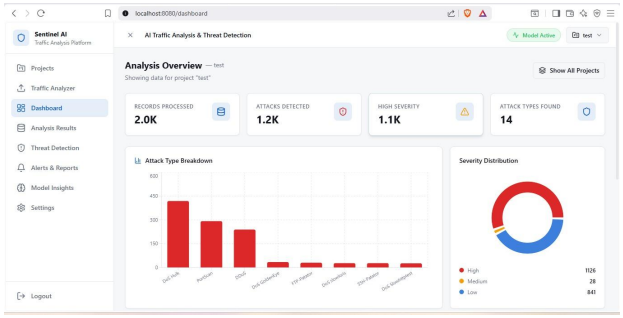

Fig. 4 Analysis overview dashboard with attack type breakdown and severity distribution

Fig. 5 depicts the Threat Detection page, which presents individual attack categories detected by the Random Forest model. Each card displays the attack type, severity label, count of detected instances, and the AI confidence score. Notably, DoS Hulk accounted for 228 detections at 99.6% confidence, PortScan yielded 141 detections at 100% confidence, and DDoS produced 119 detections at 99.9% confidence. The AI Explain button on each card invokes the LLM-based explanation module, which generates a human-readable contextual description of the detected threat. The Recent Threats Timeline at the bottom of the page logs each detection event with source IP, destination IP, and confidence score, providing analysts with granular audit information.

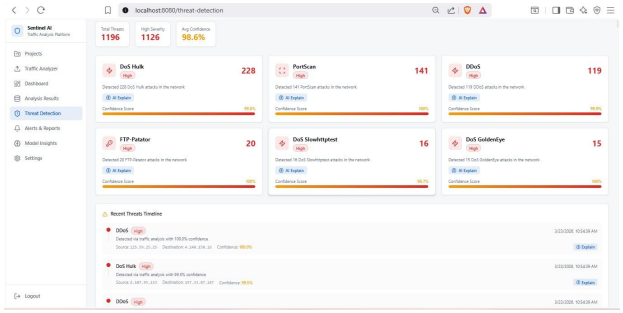

Fig. 5 Threat detection page with per-attack-type confidence scores and LLM explanation trigger

Fig. 6 shows the Analysis Results table, which lists every traffic record processed by the ML model along with its predicted attack type, confidence score, severity classification, source IP address, and destination IP address. Benign traffic records are assigned a Low severity label with no explanation required, while malicious records such as DDoS and DoS Hulk are flagged as High severity and offer an Explain button that invokes the LLM module. This granular per-record view enables analysts to trace individual attack instances and prioritise incident response based on confidence and severity.

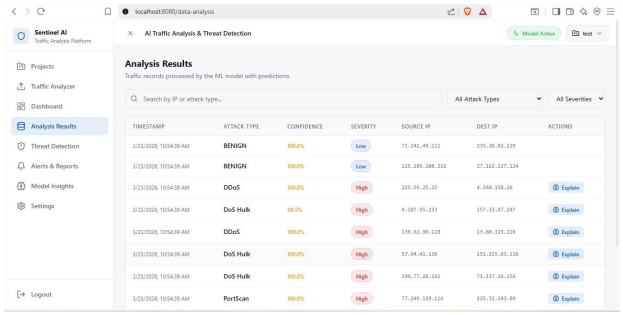

Fig. 6 Analysis results table with per-record attack classification and severity

Fig. 7 illustrates the Alerts and Reports section, which consolidates high-severity threat records with their unique Analysis IDs, detected threat types, timestamps, source IPs, confidence scores, and severity ratings. The interface

DOI: 10.48175/IJARSCT-33250

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, April 2026 Impact Factor: 8.2

supports filtering by severity level (High, Medium, Low), enabling analysts to focus on the most critical incidents. The Explain button on each alert row retrieves the LLM-generated explanation, which describes the nature of the attack, the contributing network features, the assessed risk level, and recommended mitigation actions. This integration of ML-based detection with LLM-based explanation significantly improves the interpretability and actionability of the system's output.

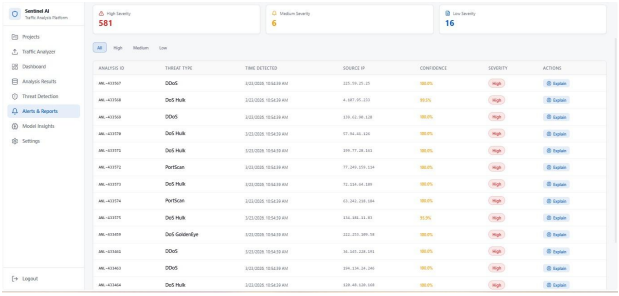

Fig. 7 Alerts and reports page with severity filtering and LLM-powered explanation

IX. CONCLUSION

An AI-driven threat detection approach for organisations to oversee their network traffic has been developed in this research. The proposed Threat Detection Model (TDM) uses Machine Learning (ML) to detect attacks and provides an explanation component that can generate summary descriptions of detected attacks.

Using the CIC-IDS2017 dataset, our experiments showed that the Random Forest classifier produced high levels of accurate identification of malicious activity in the traffic data. The anomaly detection model was used along with the supervised classifier to identify anomalous traffic patterns.

By adding a language model to the model's output, the explanation component of the model will provide an increased level of interpretability of the detection result, which will allow analysts to more easily comprehend and address potential threats.

X. FUTURE WORK

Future research will focus on:

- Implementing real-time network traffic monitoring
- Integrating advanced deep learning-based detection models
- Improving explanation quality using state-of-the-art transformer-based language models
- Deploying the system in live enterprise network environments

ACKNOWLEDGMENT

The authors would like to thank the faculty and staff of Anjalai Ammal Mahalingam Engineering College, Kovilvenni, for their support and guidance throughout this research work.

REFERENCES

[1] T. Brown et al., "Language models are few-shot learners," NeurIPS, 2020.
[2] W. Zhao et al., "A survey of large language models," ACM Computing Surveys, 2023.
[3] Y. Chen, Z. Li, and H. Zhang, "Deep learning-based network intrusion detection: Recent advances and future directions," IEEE Access, vol. 11, pp. 45612–45629, 2023.
[4] M. Naseer, A. Khan, and R. Ullah, "Explainable artificial intelligence for cybersecurity intrusion detection systems," IEEE Transactions on Artificial Intelligence, vol. 5, no. 2, pp. 210–223, 2024.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, April 2026 Impact Factor: 8.2

[5] S. Wang, J. Liu, and X. Chen, "Large language models for cybersecurity applications: Opportunities and challenges," ACM Computing Surveys, vol. 56, no. 4, 2024.
[6] R. Sharma, P. Kumar, and A. Gupta, "Hybrid machine learning approach for network intrusion detection using ensemble models," IEEE Access, vol. 12, pp. 18745–18759, 2024.
[7] H. Kim and J. Park, "AI-assisted security monitoring using explainable machine learning models," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 1123–1135, 2024.
[8] L. Zhang, Y. Wu, and M. Zhao, "Large language models for automated cyber threat analysis and incident response," IEEE Security & Privacy, vol. 22, no. 3, pp. 55–63, 2025