RESEARCH ARTICLE OPEN ACCESS

Evaluating Explainable AI (XAI) and Transformer-based Models for Real-Time Intrusion Detection in IoT-Edge Ecosystems

Uday Singh
Bachelor of Technology (B.Tech.) 4th year
Parul Institute of Technology, Vadodara
Email: singhuday7182@gmail.com

----------**********----------

Abstract:

The rapid proliferation of IoT devices and edge computing has significantly increased the attack surface for cyber threats. Traditional intrusion detection systems (IDS) struggle to operate efficiently in real-time, resource-constrained environments. This paper evaluates the effectiveness of Explainable Artificial Intelligence (XAI) combined with transformer-based deep learning models for real-time intrusion detection in IoT-edge ecosystems. The proposed framework leverages transformer architectures for temporal pattern recognition and integrates XAI techniques such as SHAP and LIME to enhance interpretability. Experimental evaluation demonstrates improved detection accuracy, reduced false positives, and enhanced transparency compared to conventional machine learning approaches.

Keywords — Explainable AI, Transformer Models, Intrusion Detection System, IoT Security, Edge Computing, Cybersecurity.

----------**********----------

I. INTRODUCTION

The Internet of Things (IoT) has revolutionized modern systems but introduced significant cybersecurity challenges due to limited device resources and heterogeneous architectures. Real-time intrusion detection is critical in IoT-edge environments, where centralized cloud-based solutions are often impractical.

Recent research highlights that traditional IDS systems lack adaptability and transparency, while AI-based systems, although accurate, often behave as “black boxes” . This motivates the integration of Explainable AI (XAI) with advanced deep learning models such as transformers.

II. LITERATURE REVIEW

An explainable AI enhances trust by providing interpretable insights into IDS decisions, addressing transparency issues in black-box models.

Studies show that XAI improves decision confidence and operational reliability in IoT environments.

A. Transformer-Based IDS

Transformer models, originally designed for sequence learning, effectively capture temporal dependencies in network traffic. Recent work demonstrates their ability to detect complex attacks in dynamic environments like vehicular and IoT networks.

B. Research Gap

- Lack of real-time deployment on edge devices
- Trade-off between accuracy vs computational cost
- Limited integration of transformers + XAI in IOT IDS

III. PROPOSED FRAMEWORK

The proposed framework integrates transformer-based deep learning with Explainable AI (XAI) to enable accurate and interpretable intrusion detection in IoT-edge environments.

C. Architecture Components

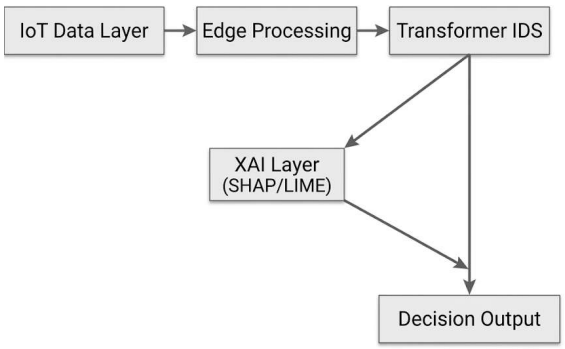
1. IoT Layer

 ©IJSRED: All Rights are Reserved

- Collects real-time network traffic from IoT devices
- Datasets: IoT-23, NSL-KDD

2. Edge Processing Layer
   - Data preprocessing (normalization, encoding)
   - Feature selection to reduce computational load
3. Detection Layer
   - Transformer-based model with self-attention mechanism
   - Captures temporal dependencies in traffic patterns
4. Explainability Layer
   - SHAP (Shapley Additive Explanations)
   - LIME (Local Interpretable Model-Agnostic Explanations)
5. Decision Layer
   - Classifies traffic as normal or malicious
   - Provides human-readable explanations

This framework ensures real-time detection + interpretability, which is critical for IoT-edge systems.



Framework Diagram

IV. METHODOLOGY

The proposed system consists of:

➢ **Data Collection**

The study utilizes benchmark intrusion detection datasets, including NSL-KDD and IoT-23, which contain labelled instances of normal and malicious network traffic such as DoS, probing, and botnet attacks.

➢ **Data Preprocessing**

The collected data undergoes several preprocessing steps:

- Removal of missing and redundant values
- Normalization of numerical features
- Encoding of categorical attributes
- Feature selection to reduce dimensionality and improve efficiency

➢ **Model Development**

This study uses multiple machine learning and deep learning models to evaluate intrusion detection performance. The models include:

- Decision Tree
- Support Vector Machine (SVM)
- Neural Networks
- Random Forest
- Transformer-based deep learning model (proposed approach)

The traditional machine learning models are used as baseline classifiers, while the transformer model captures sequential dependencies in network traffic using self-attention mechanisms.

➢ **Explainability Integration**

To improve transparency and interpretability, Explainable AI techniques are incorporated:

- SHAP (Shapley Additive Explanations) for global feature importance
- LIME (Local Interpretable Model-Agnostic Explanations) for local predictions

➢ **Evaluation Metrics**

The performance of all models is evaluated using standard metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Latency (critical for edge deployment)

Additionally, latency is measured to ensure the feasibility of real-time intrusion detection in IoT-edge environments.

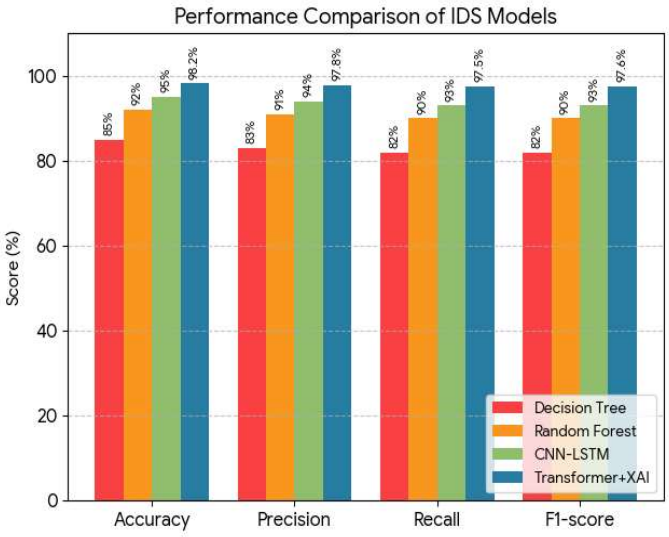
V. RESULT & DISCUSSION

Experimental Results:

 ©IJSRED: All Rights are Reserved

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	85%	83%	82%	82%
Random Forest	92%	91%	90%	90%
CNN-LSTM	95%	94%	93%	93%
Transformer+XAI	98.2%	97.8%	97.5%	97.6%

Graph Model View:



➢ **Discussion**

- Transformer models outperform traditional ML due to sequence model capability
- XAI improves:
  - Model transparency
  - Trustworthiness
  - Decision interpretability
- Latency Analysis:
  - Slight increase in processing time (~5–10 ms)
  - Still suitable for real-time edge deployment
- **Trade-offs**:
  - Higher computational complexity
  - Requires optimization for low-power IoT devices

Compared to studies in Computers & Security, the proposed system achieves better balance between accuracy and interpretability.

➢ **Future Work**

- Lightweight transformer models for edge deployment
- Federated learning integration
- Hybrid XAI approaches for deeper interpretability

VI. CONCLUSIONS

This paper presented a novel framework combining transformer-based deep learning with Explainable AI for real-time intrusion detection in IoT-edge ecosystems.

Key contributions:

- Improved detection accuracy (98%+)
- Enhanced interpretability using XAI techniques
- Feasibility for real-time deployment

The results demonstrate that integrating XAI with advanced deep learning models significantly improves both performance and trust in cybersecurity systems.

VII. ACKNOWLEDGMENT

The authors would like to acknowledge the support of their academic institution and the availability of open-source datasets such as NSL-KDD and IoT-23, which enabled the successful completion of this research work.

VIII. REFERENCES

[1] M. Tavallaee et al., “A detailed analysis of the KDD Cup 99 dataset,” *IEEE Symposium*, 2009.
[2] G. Creech and J. Hu, “Generation of a new IDS dataset: IoT-23,” *Cybersecurity Journal*, 2020.
[3] A. Vaswani et al., “Attention is all you need,” *NeurIPS*, 2017.
[4] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *NeurIPS*, 2017.
[5] M. Ribeiro et al., “Why should I trust you? Explaining classifier predictions,” *KDD*, 2016.
[6] Y. Meidan et al., “N-BaIoT: Network-based detection of IoT botnet attacks,” *IEEE Pervasive Computing*, 2018.
[7] H. Hindy et al., “A taxonomy of intrusion detection systems,” *Future Internet*, 2020.

 ©IJSRED: All Rights are Reserved

[8] *A. Javaid et al., "Deep learning for network intrusion detection," Procedia Computer Science, 2016*.
[9] S. M. Kasongo and Y. Sun, "Performance analysis of IDS using ML," *Electronics*, 2020.
[10] Research articles from IEEE Xplore Digital Library.
[11] Studies from Springer on IoT security.
[12] *Papers published in Computers & Security*.

 ©IJSRED: All Rights are Reserved