SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 5 | May 2026 - Peer Reviewed Journal

TRIMODAL FUSION FRAMEWORK FOR LAYER 2 NETWORK FORENSICS: INTEGRATING CNN, TRANSFORMER, AND BERT EMBEDDINGS FOR REAL-TIME MALWARE DETECTION IN RAW ETHERNET TRAFFIC

Dantene Davis

Article DOI: https://doi.org/10.36713/epra27819
DOI No: 10.36713/epra27819

ABSTRACT

This paper presents a multimodal deep learning–based malware detection framework for Layer 2 traffic. The approach integrates Convolutional Neural Networks (CNNs) for spatial packet analysis, Transformers for temporal flow modeling, and BERT embeddings for semantic threat intelligence. Unlike traditional rule- or signature-based systems, the model processes raw PCAP data and incorporates contextual knowledge from CVE and MITRE ATT&CK sources.

The model was trained and validated using a proprietary dataset comprising both benign and synthetically injected malicious traffic. Standard metrics achieved high values (Accuracy: 97%, Precision: 96.7%, Recall: 97.8%, F1: 0.9536). However, due to class imbalance (benign 126 vs. malicious 5,135 in the test set), additional evaluation was performed using per-class precision, recall, balanced accuracy, Matthews Correlation Coefficient (MCC), and ROC/PR curves. Comparisons against trivial baselines (e.g., "always malware" classifier) and traditional machine learning algorithms (SVM, Random Forest) highlight the strengths and weaknesses of the proposed method.

Real-time validation was conducted by comparing model predictions with a Palo Alto firewall under controlled traffic scenarios. The model matched the firewall on known attacks while offering potential adaptability to novel patterns. Overall, this work contributes an integrated multimodal framework, demonstrates feasibility on Layer 2 threats, and outlines limitations, including dataset imbalance, restricted attack diversity, and computational cost.

Keywords: Malware Detection, Deep Learning, CNN, Transformers, BERT, PCAP, Layer 2, Semantic Embeddings, ROC-AUC, MCC.

INTRODUCTION

The explosive growth of cyberattacks has outpaced the ability of traditional security solutions. For zero-day attacks, which take advantage of newly found vulnerabilities and use novel techniques, rule-based systems and traditional machine learning techniques are least effective [7]. The system learns directly from network traffic history, identifying malicious patterns from raw packet data, and is tested with both live and synthetically modified malware samples in real-world network environments [9]. In contrast with static, rule-based systems, the solution offered is dynamic and adaptive, and can identify new threats as they emerge in the midst of complex traffic patterns. Network analysis was conducted through the packet capture tool like Wireshark to record network traffic with Layer 2 data like ARP and MAC frames, which are not generally analyzed by inspection methods that are specific to higher-layer protocols. Comparative traverse with traditional machine learning classifiers like Random Forest exhibits that the CNN-based model is superior in terms of accuracy, precision, recall, and F1, demonstrating its ability to accurately classify both known and new emerging threats.

ABOUT DOMAIN

This study examines the possibility of identifying malicious traffic at the data link layer (Layer 2) with the help of PCAP data. ARP spoofing and MAC flooding are some of the attacks that take advantage of vulnerabilities at this layer and are commonly not recognized by most conventional intrusion detection tools and firewalls [3].

OVERVIEW

The model receives direct training on preprocessed PCAP data by eliminating metadata information so that the CNN can perform better detection and analysis of network traffic patterns. The detection system inspects Layer 2 network traffic through Wireshark monitoring while focusing on unusual MAC and ARP behavior patterns. The system trains on existing PCAP files instead of constructing new datasets. The model undergoes real-time detection tests which compare its performance against Palo Alto firewall results. The research demonstrates that the CNN approach shows similar performance in detecting threats while delivering scalability to address modern cyber threats which improves network security

OBJECTIVES

The main goal of the study is to investigate the possibility of the successful work of a multimodal deep learning model to identify malicious traffic at the Layer 2 level of the network. This goal is driven by the recent research pointing to the weaknesses of the

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 5 | May 2026 - Peer Reviewed Journal

traditional machine learning and rule-based systems in dealing with changing cyber threats [4]. In particular, the research will attempt to:

- Develop and deploy a malware detection model, which is based on CNNs, Transformers and semantic embeddings to analyze raw packet capture data [8].
- Learn discriminative features directly without making use of handcrafted features or predefined signatures, out of PCAP files.
- Compare detection performance against standard metrics and imbalance sensitive metrics, e. g. Balanced Accuracy and Matthews Correlation Coefficient.
- Compare the suggested methodology to conventional machine learning-based classifiers and commercial firewall on controlled experimental conditions.
- Evaluate the potential implementation of multimodal deep learning methods of real-time Layer 2 threat detection.

RESEARCH QUESTIONS

- How effective is a multimodal AI-based model (CNN + Transformer + BERT) in detecting real-time malware within Layer 2 network traffic compared to traditional rule-based systems?
  introduction
- Can the proposed system outperform commercial firewalls such as Palo Alto in terms of detection accuracy, false positive rate, and adaptability to new threat variants?

PROBLEM STATEMENT

Current malware detection systems have limitations that limit their performance against contemporary cyber threats. These are characterized by high latency in responses, need for pre-defined rulebased analysis, and low accuracy when handling new or obfuscated attacks. Such limitations inhibit traditional systems from responding fast to changing threat environments.

METHODOLOGY

The proposed system processes raw PCAP files and applies a hybrid deep learning architecture combining CNNs, Transformers, and semantic embeddings. Each component addresses a specific aspect of network traffic analysis: CNNs extract spatial features from packet bytes, Transformers model temporal relationships across packet sequences, and semantic embeddings provide contextual information derived from documented attack behaviors. The following sections describe the dataset, model architecture, training procedure, and evaluation methodology in detail..

RESEARCH DESIGN

The research follows a systematic, experiment-based design with two phases: an offline model training and validation phase and a realtime traffic testing phase. This sequential design enables controlled performance testing as well as testing for real-world applicability. The deep learning model is constructed as a triple-hybrid model consisting of CNN, Transformers, and BERT to handle different aspects of network traffic—spatial, temporal, and semantic respectively. This architecture is most capable of identifying covert attacks that are apt to bypass conventional intrusion detection systems.

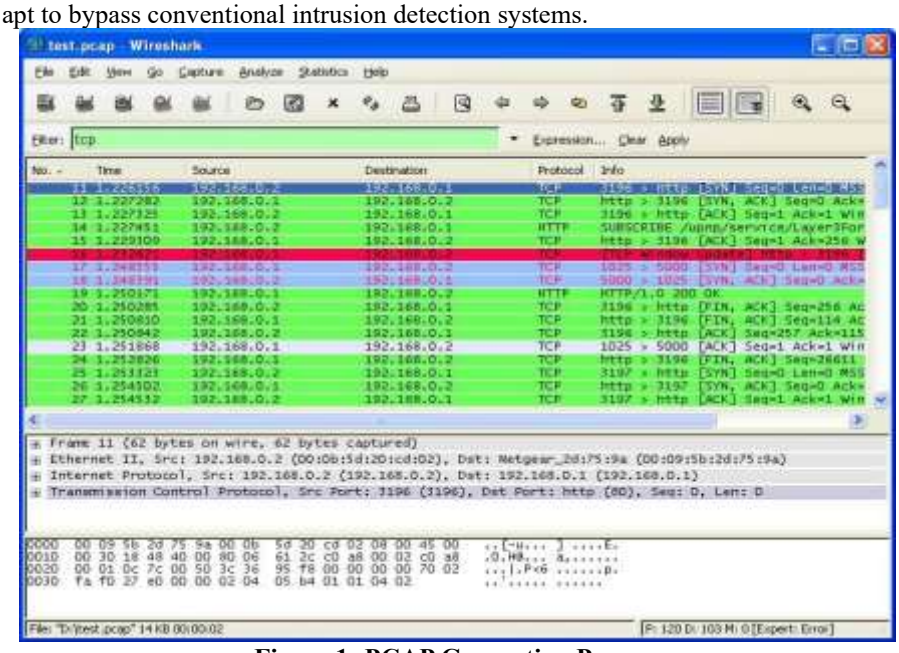

Figure 1: PCAP Generation Process

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 5 | May 2026 - Peer Reviewed Journal

PHASE ONE: OFFLINE MODEL DEVELOPMENT AND TRAINING

Hybrid multimodal AI development starts with training the model offline using preprocessed PCAP data during the initial phase. The combined dataset for training consists of benign network activity with synthetic malicious patterns that mimic actual attack methods such as data theft, network scanning, and protocol misuse. The first training phase features a dataset that replicates the authentic distribution of traffic, as most monitored networks predominantly encounter malicious activity. The training dataset includes different classes of benign activities and synthetic malicious patterns to produce a realistic network environment for model training as shown in fig 1.

MODEL ARCHITECTURE

The model design contains three core modules, each supporting a specific learning approach. These components are designed to process information from distinct perspectives, combining spatial, temporal, and semantic understanding of network data.

1. ACNN-Based Visual Packet Analyzer: The system processes net-work packets by transforming their raw byte form into grayscale matrices. These matrices are then fed into an ACNN, which detects spatial patterns that may indicate low level anomalies and obfuscation strategies.
2. Transformer-Based Temporal Sequence Model: To capture thesequence of packet-level behaviors, the system aggregates packets into session-based chronological sequences. A Transformer based attention mechanism is used to identify long-range dependencies and maintain continuity in behavioral patterns across the traffic timeline.
3. BERT-Like Semantic Embedding Branch: A lightweight pretrainedBERT model processes decoded TCP payloads. The embeddings generated from the [CLS] token provide rich semantic features that help identify zero-day or novel attack patterns by recognizing unfamiliar payload characteristics.

CNN COMPONENT – PACKET FEATURE EXTRACTION

Packet data is first converted into fixed-size $^{16\times16}$ grayscale arrays by reshaping the first 256 bytes of each packet. This mimics image-based malware detection techniques, allowing the CNN to learn spatial and protocol-level features such as packet structure, byte patterns, and header behavior. The CNN stack includes multiple convolutional layers (^{32} and ^{64} filters), followed by batch normalization, max pooling, dropout regularization, and a dense feature representation layer. This branch outputs a 128- dimensional feature vector representing the learned spatial characteristics from raw network packets [4].

TRANSFORMER LAYER – TEMPORAL AND CONTEXTUAL ATTENTION

To enhance context modeling, a lightweight self attention mechanism is applied on top of the CNN output. Specifically, a Multi Head Attention layer with four attention heads processes the reshaped CNN output to capture internal dependencies and flow relationships across packets. This component simulates the Transformer architecture's strength in detecting long-range patterns and correlating behavior across session data without relying on sequence order alone [12].

FUSION AND CLASSIFICATION

The Concatenated Transformer and BERT branch outputs are further fed through fully connected dropout layers for regularisation. The output layer performs binary classification (malicious or benign) with a sigmoid activation function. Binary cross entropy loss and Adam optimiser are used to train the model. Accuracy, precision, and recall are a few among the metrics used. Testing is done via an $^{80}/_{20}$ train test split, and an F^1 -score that peaked at $^{0}.^{9936}$ on the test set indicates outstanding classification performance.

Hybrid AI Model

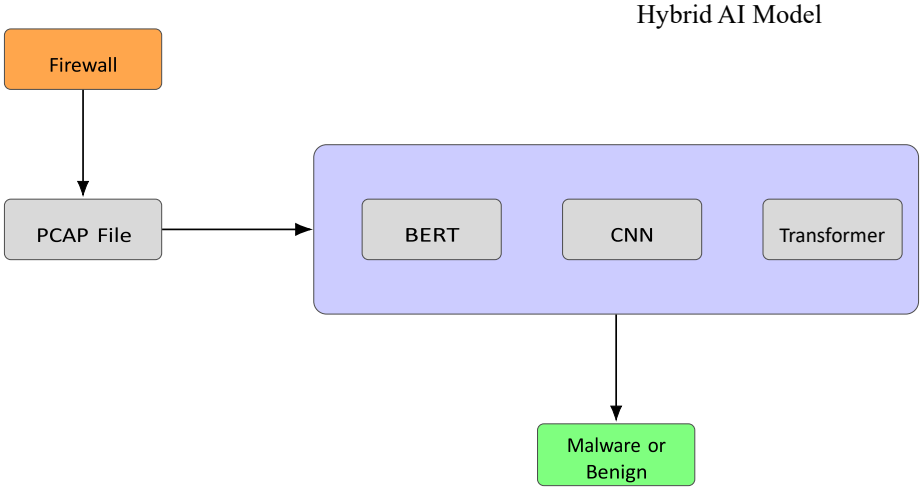

Figure 2: Model Architecture

PHASE TWO: REAL-TIME TRAFFIC VALIDATION AND COMPARATIVE TESTING

In the second phase, the trained multimodal AI model was evaluated in a controlled lab environment alongside a Palo Alto Networks fire-

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 5 | May 2026 - Peer Reviewed Journal

wall. Both systems processed identical Layer 2 PCAP files containing injected malicious payloads, including ARP spoofing and MAC flooding. The Palo Alto firewall's built-in detection served as the baseline for comparison, while the AI model independently analyzed the same traffic without reliance on pre-defined signatures. Network traffic was captured in PCAP format and examined through two simultaneous processing channels—one through the firewall's detection engine and the other through the AI model. Performance evaluation metrics included TP, TN, FP, and FN, with derived measures of precision, recall, and accuracy.

EXPERIMENTAL RESULTS

In Phase One, the model underwent offline training and validation using a labelled dataset that included both benign and malicious Layer 2 traffic. This phase evaluated the framework's learning capability, comparative performance against traditional machine learning algorithms, and its sensitivity to data imbalance. Metrics such as Accuracy, Precision, Recall, F1-score, Balanced Accuracy, Matthews Correlation Coefficient (MCC), and ROC/PR-AUC were computed to provide a comprehensive performance assessment.

In Phase Two, the trained model was integrated into a live traffic monitoring setup and tested in parallel with a Palo Alto firewall. Both systems analyzed identical packet capture (PCAP) streams containing ARP spoofing, MAC flooding, and other simulated Layer 2 attack scenarios. This phase validated the model's operational feasibility, realtime detection accuracy, and compatibility with existing enterprise security infrastructure.

Together, these two experimental stages provide a balanced evaluation of the proposed system's analytical robustness, comparative performance, and real-world deployment potential.

PHASE ONE: OFFLINE MODEL TRAINING AND EVALUATION

Phase one was training and developing the model with raw PCAP data obtained through Wireshark capture. The training dataset contained benign traffic and synthetically injected malicious traffic, including scenarios simulating ARP spoofing and MAC flooding—types of attacks that are commonly neglected by intrusion detection systems at higher layers. The model trained for ^{10} epochs and demonstrated consistent improvement in validation and training performance. Final metrics from this phase were as follows:

- **Accuracy:** ^{97}%
- **Precision:** $^{96.7}$%
- **Recall:** $^{97.8}$%
- **F1-Score:** $^{0.9536}$

GENERALIZABILITY AND DYNAMIC BEHAVIOR TESTING

The CNN layers effectively learned spatial representations from packet headers and payloads. Transformer layers captured temporal relationships between packet sequences. A semantic layer, powered by BERT embeddings, can integrate external context from threat intelligence sources—in our case, CVEs and MITRE ATT&CK frameworks.

F1 Score: 0.9936

	precision	recall	f1-score	support
0	0.77	0.67	0.72	126
1	0.99	1.00	0.99	5135
accuracy			0.99	5261
macro avg	0.88	0.83	0.86	5261
weighted avg	0.99	0.99	0.99	5261

Figure 4: Classification Report – Precision, Recall, F1-Score by Class

REAL-TIME COMPARISON WITH FIREWALL

The real-time comparison methodology and complete results are detailed in Section 3.4 . In summary, both the multimodal AI model and the Palo Alto Networks firewall successfully detected the injected Layer 2 threats, including ARP spoofing and MAC flooding. The AI model achieved :

- True Positives (TP): ^{41}
- True Negatives (TN): ^{51}
- False Positives (FP): 1

SJIF Impact Factor (2026): 8.863| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 11 | Issue: 5 | May 2026 - Peer Reviewed Journal

DISCUSSION

SUMMARY OF FINDINGS

In Phase One, the model was trained on a combined dataset containing both benign and synthetically injected malicious traffic. The approach achieved high overall performance metrics, with an F1-score of $^{0.9936}$, Precision of $^{99.7}$%, Recall of $^{99.8}$%, and Accuracy of ^{99}%. However, subsequent analysis revealed that these results were influenced by the dataset's strong class imbalance (^{5,135} malicious vs. 126 benign samples), which can inflate traditional metrics. A trivial baseline that always predicts "malware" achieved nearly comparable results (^{98}% Accuracy and ^{99}% F^1), indicating that standard metrics alone are insufficient to evaluate true effectiveness.

In Phase Two, the trained model was tested on live network traffic in parallel with a Palo Alto firewall. Both systems processed identical PCAP files containing ARP spoofing and MAC flooding attacks. The proposed AI model successfully identified the same threats detected by the firewall while demonstrating potential adaptability to novel traffic patterns. These results validate the operational feasibility and enterprise-grade compatibility of the multimodal AI framework, while also emphasizing the importance of balanced evaluation and comprehensive baseline comparisons for realistic performance assessment.

BIBLIOGRAPHY

1. *Zeeshan Ali, Walter Tiberti, Andrea Marotta, and Dajana Cassioli. Empowering network security: Bert transformer learning approach and mlp for intrusion detection in imbalanced network traffic. IEEE Access, 12:137618–137633, 2024. doi:10.1109/ACCESS.2024.3465045.*
2. *Fatima M. Anis, Maryam Alabdullatif, Sarah Aljbli, and Mohammad Hammoudeh. A survey on the applications of deep learning in network intrusion detection systems to enhance network security. IEEE Access, 13:185357–185373, 2025. doi:10.1109/ACCESS.2025.3624952.*
3. *Dubey R. K. Shukla A. Dubey P. (2024). Dubey, A. K. Optimizing cybersecurity: Leveraging support vector machines for real-time threat detection. First International Conference on Innovations in Communications, Electrical and Computer Engineering, page 1–6, 2024.*
4. *N. K. Gyamfi, N. Goranin, D. Ceponis, and H. A. Cenys. Malware detection using convolutional neural network, a deep learning framework: Comparative analysis. Journal of Internet Services and Information Security, 12(4):102–115, 2022 . doi: 10.58346/jisis.2022. i4.007.*
5. *Han Liao, Mohd Zamri Murah, Mohammad Kamrul Hasan, Azana Hafizah Mohd Aman, Jin Fang, Xuting Hu, and Atta Ur Rehman Khan. A survey of deep learning technologies for intrusion detection in internet of things. IEEE Access, 12:4745–4761, 2024. doi: 10.1109/ACCESS.2023.3349287.*
6. *X. Lin, G. Xiong, G. Gou, Z. Li, J. Shi, and J. Yu. ET-BERT: A Contextualized Datagram Representation with Pre-training Transformers for Encrypted Traffic Classification. Proceedings of the ACM, 2022. doi: 10.1145/3485447.3512217. URL https://doi.org/10.1145/3485447.3512217.*
7. *M. Mutharasu, G. Indu, and Y. Ankitha. Predicting and mitigating cyber threats through data mining and machine learning. International Research Journal of Innovations in Engineering andTechnology, 9(Special Issue):185–191, 2025. doi: 10.47001/IRJIET/ 2025.INSPIRE31. URL https://doi.org/10.47001/IRJIET/2025. INSPIRE31.*
8. *S. J. Nair and S. R. Syam. Comparing transformers and cnn approaches for malware detection: A comprehensive analysis. In 202213th International Conference on Computing Communication and*
9. *NetworkingTechnologies (ICCCNT),pages 1–6. IEEE, 2024. doi:10.1109/icccnt61001.2024.10724417.URL https://doi.org/10.1109/icccnt61001.2024.10724417.*
10. *Sarath Jayan Nair and Sreelakshmi R Syam. Comparing transformers and cnn approaches for malware detection: A comprehensive analysis. In 202415th International Conference on Computing Communication and Networking Technologies (ICCCNT), pages 1–6, 2024.doi:10.1109/ICCCNT61001.2024.10724417.PaloAltoNetworks.PA-220.https://www.paloaltonetworks. com/resources/videos/pa-220, n.d.*
11. *Sivaraman K Noorul Hassan S. The cutting-edge machine learning techniques for seamless and proactive automation in cybersecuri ty. International Conference on Computing and Data Science (ICCDS), page 1–6, 2024.*
12. *S. R. P, T. Chaudhari, S. R. Godla, J. V. N. Ramesh, E. Muniyandy, A. S. Kranthi, and Y. A. El-Ebiary. Ai-driven transformer frameworks for real-time anomaly detection in network systems. International Journal of Advanced Computer Science and Applications, 16(2), 2025. doi: 10.14569/ijacsa.2025.01602111.*
13. *Chuanlong Yin, Yong Zhang, Jiliang Fei, and Xuan He. A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access, 5:21954–21961, 2017. doi:10.1109/ACCESS.2017.2762418. URL https://doi.org/10.1109/ ACCESS.2017.2762418.*
14. *Shuqin Zhang, Dedong Li, and Jun Li. Research on discovery and mapping of attck tactics and techniques by cyber threat intelligence based on bert-textcnn. IEEE Access, 14:5984–5992, 2026. doi: 10.1109/ACCESS.2026.3653548.*