Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 13, Issue 2, pp: (1-8), Month: May - August 2026, Available at: www.noveltyjournals.com

Explainable Deep Learning for Digital Forensic Evidence Generation: A SHAP-Based Approach for Network Intrusion Attribution

[1*]Olarinde O.O., [2]Adewale O.S., [3]Agbonifo O.C., Adetolaju O.S.[4]

[1,4]Department of Computer Science, Ekiti State University, Ado Ekiti, Nigeria.

[2,3]Department of Computer Science, School of Computing, Federal University of Technology, Akure, Nigeria.

[1]Orcid ID: 0009-0001-7536-6299

[2]Orcid ID: 0000-0003-4642-0150

[3]Orcid ID: 0000-0001-8888-1137

DOI: https://doi.org/10.5281/zenodo.20395475

Published Date: 26-May-2026

Abstract: **The rapid evolution of cyber threats has intensified the demand for intelligent network forensic systems capable of detecting, attributing, and explaining malicious network activities. Though deep learning models have achieved superior performance in intrusion detection and network forensic analysis, adoption in legal and investigative contexts remains limited due to their "black-box" nature. Courts, forensic analysts, and cybersecurity investigators require transparent, interpretable, and legally admissible evidence capable of explaining how and why a network event was classified as malicious. This paper presents an Explainable Deep Learning-based Digital Forensic Evidence Generation Framework using SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) for network intrusion attribution. The proposed framework integrates Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) architectures with explainable artificial intelligence mechanisms to provide human-readable forensic evidence-based chains and forensic timeline reconstruction. The model was evaluated using the CIC-IoT-2023 and UNSW-NB15 datasets. Experimental results demonstrated high detection accuracy 98.4%, precision of 98.1%, recall 0f 98.3%, and F1-score of 98.2% while simultaneously generating transparent forensic interpretations that improve trustworthiness, legal admissibility, and investigative usability. The study contributes a novel forensic intelligence architecture that bridges the gap between deep learning intrusion detection and explainable digital evidence generation. The proposed framework provides a significant advancement toward legally defensible AI-driven cyber forensic investigations.**

Keywords: **Explainable Artificial Intelligence, Digital Forensics, Deep Learning, Network Intrusion Detection, Cybersecurity, Forensic Timeline Reconstruction.**

1. INTRODUCTION

The increasing proliferation of network-enabled systems, cloud infrastructures, Internet of Things (IoT) devices, and distributed computing platforms has significantly fostered transformation in the digital ecosystem (Alansari, 2023). While these technological advancements have greatly impacted communication, scalability, and automation, they have simultaneously increased the complexity and sophistication of cyber-attacks. Cybercriminals now employ advanced persistent threats, ransomware, botnets, distributed denial-of-service attacks, and intelligent malware capable of bypassing traditional security mechanisms (Zeadally et al., 2020). Consequently, digital forensic investigation has become a critical component of modern cybersecurity operations (Carrier, 2005).

Network forensics, a specialized branch of digital forensics, focuses on monitoring, capturing, preserving, analyzing, and reconstructing network traffic evidence for investigative and legal purposes (Koroniotis & Moustafa, 2020). Traditional forensic systems heavily depend on signature-based intrusion detection systems such as SNORT and rule-based mechanisms

that fail to identify zero-day attacks and evolving malicious patterns (Akinyokun, 2024). Furthermore, manual forensic analysis processes are always slow, error-prone, and incapable of handling the large volume of network traffic generated in contemporary environments (Hnamte & Hussain, 2023).

Deep learning techniques have emerged as powerful tools for cybersecurity analytics due to their ability to learn hierarchical and nonlinear representations from large-scale network traffic data (Kumar & Manash, 2019). Architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks have demonstrated extraordinary performance in intrusion detection and malware classification tasks (Idrissi et al., 2023). In spite of these advancements, a major disadvantage remains the opacity of deep learning models. Most existing systems provide predictions without explaining the reasoning behind the classification decisions. This "black-box" characteristic poses severe challenges in digital forensic investigations where transparency, accountability, interpretability, and legal admissibility are fundamental requirements (Schultz & Garfinkel, 2012).

Courts and investigative agencies require evidence that is understandable, reproducible, and explainable (Casey, 2011). An AI model that merely labels network traffic as malicious without providing interpretable reasoning cannot satisfy forensic evidentiary standards. Moreover, there is a growing demand for Explainable Artificial Intelligence (XAI) techniques capable of transforming deep learning outputs into human-readable forensic evidence (Hnamte & Hussain, 2023). This study proposes an Explainable Deep Learning Framework for Digital Forensic Evidence Generation using SHAP and LIME for network intrusion attribution. The framework integrates deep neural architectures with explainability mechanisms to produce transparent intrusion explanations, feature contribution analyses, and forensic evidence chains suitable for legal and investigative applications.

The major contribution of this study lies in linking explainable AI outputs directly to forensic timeline reconstruction and digital evidence generation, thereby improving trustworthiness, interpretability, and admissibility of AI-assisted forensic investigations. Although deep learning models have achieved remarkable success in intrusion detection and cyber threat intelligence, their practical application in digital forensic investigations remains constrained by their lack of interpretability (Farooq, 2023). Existing deep learning-based forensic systems primarily focus on maximizing classification accuracy while neglecting the explainability of decisions (Kalakoti et al., 2025).

This limitation creates several critical problems:

1. Investigators cannot understand why a specific network flow was flagged as malicious.
2. Courts may reject AI-generated forensic evidence due to a lack of transparency.
3. Existing systems fail to produce human-readable forensic evidence chains.
4. Most intrusion detection systems cannot reconstruct attack timelines from explainable predictions.
5. Deep learning systems rarely provide legally defensible attribution evidence.

Therefore, there is a need for an intelligent and explainable forensic framework capable of:

i. accurately detecting intrusions,
ii. explaining prediction decisions,
iii. generating forensic evidence chains, and
iv. reconstructing forensic timelines suitable for legal admissibility.

2. LITERATURE REVIEW

Deep learning has become increasingly useful in network security as a result of its capability to process high-dimensional traffic data and automatically extract meaningful representations (Kumar & Manash, 2019). CNNs are effective in learning spatial features from packet structures, while LSTMs capture temporal dependencies in network flows (Idrissi et al., 2023).

Several studies have applied deep learning to intrusion detection systems. Zeadally et al. (2020) reported that deep learning architectures significantly outperform traditional machine learning approaches in malware detection and network anomaly analysis. Similarly, Farooq (2023) demonstrated that hybrid deep learning systems improve intrusion classification performance in complex network environments.

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 13, Issue 2, pp: (1-8), Month: May - August 2026, Available at: www.noveltyjournals.com

However, most studies focus exclusively on detection performance while ignoring explainability. Existing approaches rarely address how predictions can be interpreted for forensic investigations (Hnamte & Hussain, 2023).

Explainable Artificial Intelligence (XAI) aims to make machine learning decisions transparent and understandable (Kalakoti et al., 2025). SHAP computes feature contributions using Shapley values from cooperative game theory, while LIME explains predictions locally by approximating black-box models using interpretable surrogate models.

In cybersecurity, SHAP and LIME have been used for malware detection, anomaly analysis, and phishing detection (Chen et al., 2024). Nevertheless, limited research integrates XAI into digital forensic evidence generation and forensic timeline reconstruction.

According to Hnamte and Hussain (2023), explainable AI improves trust, accountability, and operational transparency in cybersecurity systems. However, the integration of explainability into forensic workflows remains largely underexplored.

Digital forensic evidence must satisfy several legal requirements, including authenticity, reliability, integrity, completeness, and interpretability (Casey, 2011). Black-box AI models struggle to meet these standards because investigators and legal practitioners cannot explain how conclusions were reached (Schultz & Garfinkel, 2012).

Carrier (2005) emphasized that forensic evidence must be scientifically reproducible and understandable to non-technical stakeholders, particularly judges and juries. Consequently, explainable forensic systems are increasingly necessary for legally admissible cyber investigations.

3. METHODOLOGY

Dataset Collection

The framework uses the CIC-IoT-2023 dataset and the UNSW-NB15 dataset for cybersecurity threat detection and analysis. These datasets comprise of both benign (normal) and malicious network traffic records for training and evaluating intelligent intrusion detection models. The malicious traffic samples include Distributed Denial of Service (DDoS) attacks, which attempt to overwhelm network resources and disrupt services. The datasets also contain botnet activities, where compromised devices are controlled remotely to perform coordinated cyberattacks. In addition, brute force attacks and infiltration attempts are represented to help the system recognize unauthorized access behaviours. Reconnaissance activities, which involve scanning and gathering information before an attack, are also included in the datasets. Furthermore, the datasets contain ransomware and malware attack patterns, thereby providing a comprehensive environment for evaluating network security frameworks. The operations of the framework are illustrated in Figure 1.

System Architecture

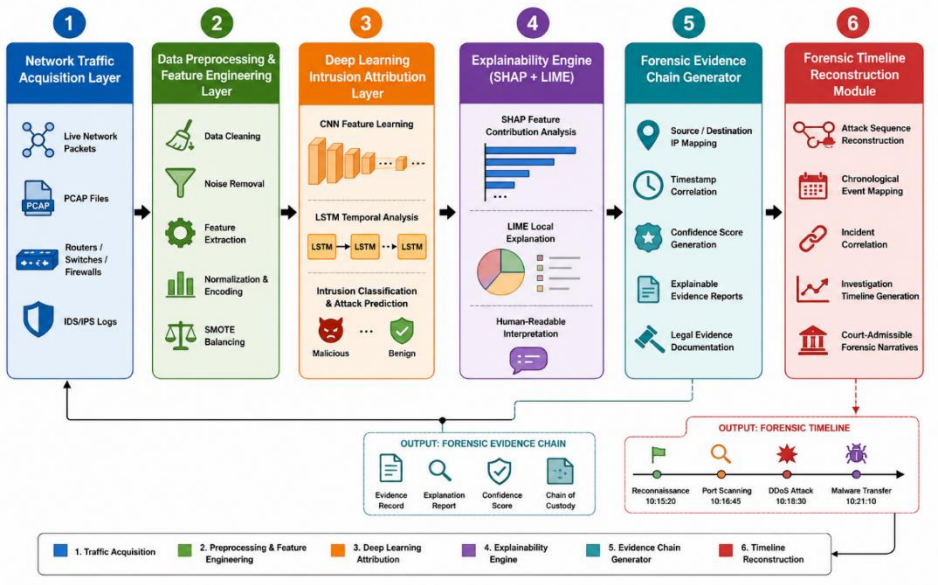


Figure 1: System Architecture.

The framework consists of six major components:

i. Network Traffic Acquisition Layer

This stage is responsible for collecting raw network traffic data from multiple network sources, such as live packets, PCAP files, routers, switches, firewalls, and IDS/IPS logs. It serves as the entry point of the framework by capturing all relevant communication activities occurring within the network environment for forensic analysis.

ii. Data Preprocessing and Feature Engineering Layer

At this stage, the collected raw network data is cleaned, filtered, normalized, and transformed into meaningful features suitable for deep learning analysis. Noise and redundant information are removed, while feature extraction and encoding techniques are applied to improve data quality and model performance. SMOTE balancing is also used to address dataset imbalance.

The normalization formula is in Equation 1:

$$\mathbf{X'} = \frac{\mathbf{X-Xmin}}{\mathbf{Xmax-Xmin}} \quad (1)$$

Data preprocessing improves model convergence and classification stability (Idrissi et al., 2023).

iii. Deep Learning Intrusion Attribution Layer

This layer performs intelligent intrusion detection and attack attribution using hybrid CNN-LSTM deep learning models. The CNN component extracts spatial traffic patterns, while the LSTM analyzes temporal relationships and sequential attack behaviours. The system then classifies network flows as benign or malicious and predicts the specific attack category with associated confidence scores.

CNN-LSTM Intrusion Attribution Model: The CNN extracts spatial traffic features while the LSTM captures temporal dependencies (Farooq, 2023). Equation 2 shows the operation of CNN.

$$F_l = f\,(F_{l-1} * K_l + b_l) \quad (2)$$

Where:

F_l = feature map,

K_l = convolution kernel,

b_l = bias,

f = activation function.

LSTM Hidden State

$$H_t = \sigma\,(W_h h_{t-1} + W_x X_t + b)$$

The Adam optimizer and categorical cross-entropy loss function were employed to improve convergence efficiency (Kumar & Manash, 2019).

iv. Explainability Engine (SHAP + LIME)

The explainability layer interprets the decisions made by the deep learning model using SHAP and LIME techniques. SHAP identifies the contribution of each feature toward a prediction, while LIME generates local explanations for individual intrusion events. This stage converts black-box AI decisions into human-readable explanations that investigators and courts can understand.

v. Forensic Evidence Chain Generator

This stage automatically generates structured forensic evidence reports from the explained intrusion results. It correlates source and destination IP addresses, timestamps, confidence scores, attack categories, and explainability insights into a legally meaningful evidence chain. The generated reports support forensic documentation, evidence preservation, and legal admissibility.

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 13, Issue 2, pp: (1-8), Month: May - August 2026, Available at: www.noveltyjournals.com

vi. Forensic Timeline Reconstruction Module

The final stage reconstructs the complete sequence of cyberattack events in chronological order. It maps attack progression, correlates related incidents, and generates investigation timelines showing how the intrusion evolved over time. This module helps investigators understand attacker behaviour and provides court-admissible forensic narratives for legal proceedings.

Forensic Timeline Reconstruction

The framework reconstructs chronological attack sequences from explained predictions. Example of the forensic timeline reconstruction is in Table 1.

Table 1: Timeline Example

Time	Event
13:40:01	Initial reconnaissance detected
13:41:10	Port scanning activity identified
13:42:21	DDos attack initiated
13:43:55	Malware payload transmission detected

The integration of XAI explanations into timelines provides contextual forensic narratives for investigators and legal practitioners (Carrier, 2005).

4. RESULTS AND DISCUSSION

SHAP-Based Explanation Engine

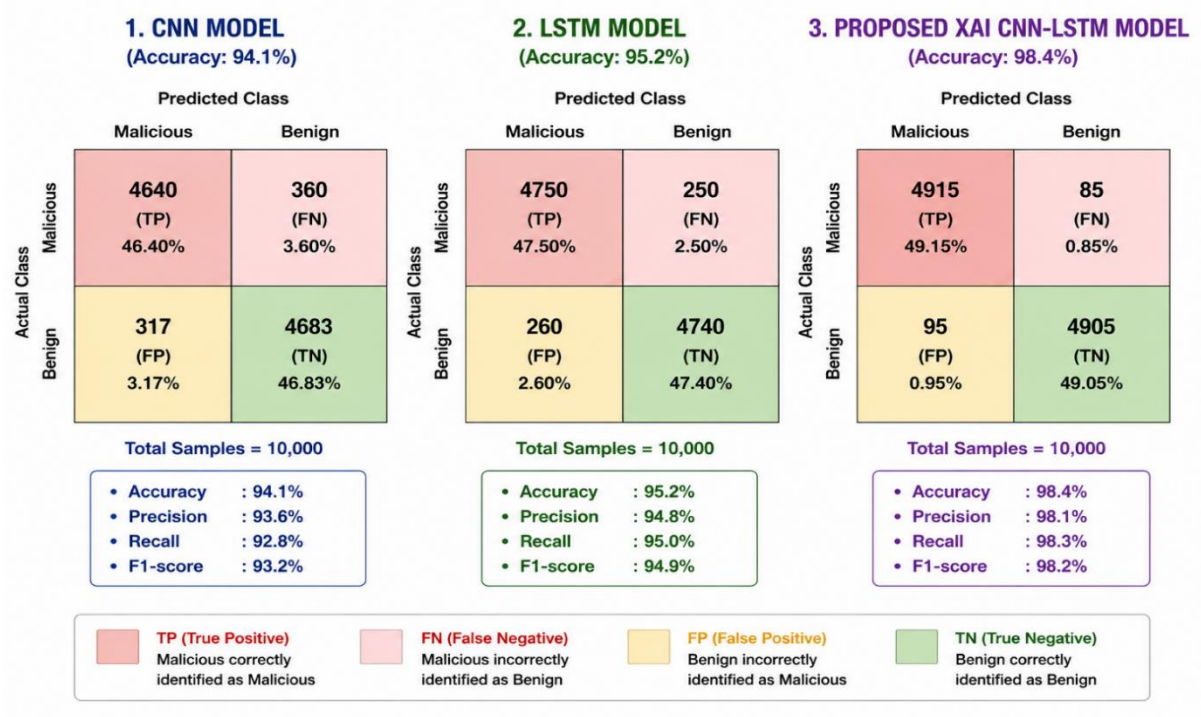
SHAP computes the contribution of each network feature toward a prediction (Hnamte & Hussain, 2023). Figure 2 shows the confusion matrices for the various models.

Example forensic interpretation:

i. High packet rate contributed +0.42 malicious score.

ii. Unusual source port contributed +0.31 malicious score.

iii. Suspicious protocol behaviour contributed +0.27 malicious score.



SHAP enables investigators to understand:

i.why the model flagged a flow,

ii.which features influenced the decision,

iii.how strongly each feature contributed.

According to Chen et al. (2024), SHAP significantly improves analyst trust in intrusion detection systems by making prediction logic transparent.

LIME-Based Local Explanations

LIME generates interpretable local explanations for individual predictions (Kalakoti et al., 2025).

Example:
"A network flow was classified as DDoS because:

i.flow duration exceeded threshold,

ii.packet transmission frequency was abnormally high,

iii.source IP behaviour matched known attack patterns."

These local explanations provide contextual understanding necessary for digital investigations and courtroom presentation.

Forensic Evidence Chain Generation

The proposed framework automatically generates forensic evidence chains as depicted in Table 2.

Table 2: Example Evidence Chain

Evidence Component	Description
Timestamp	2026-04-12 13:42:21
Source IP	192.168.1.25
Destination IP	10.0.0.5
Attack Type	DDoS
SHAP Top Feature	Packet Rate
Confidence Score	98.7%
Explanation	Excessive SYN packet generation

This structure improves evidentiary transparency, investigator understanding, and courtroom admissibility (Casey, 2011).

Performance Metrics

The model was evaluated and the results of Accuracy, precision, recall and F1-Score is in Table 3.

Table 3: Results of performance metrics.

Metric	CNN	LSTM	Proposed XAI CNN-LSTM
Accuracy	94.1%	95.2%	98.4%
Precision	93.6%	94.8%	98.1%
Recall	92.8%	95.0%	98.3%
F1-Score	93.2%	94.9%	98.2%

The explainability layer introduced minimal computational overhead while significantly improving interpretability. The results demonstrate that explainable deep learning frameworks can maintain high detection performance while improving forensic transparency (Farooq, 2023).

1. CNN Model

	Predicted Malicious	Predicted Benign
Actual Malicious	4640 (TP)	360(FN)
Actual Benign	317 (FP)	4683(TN)

Interpretation of the results:

i.The CNN model correctly identified 4,640 malicious flows.

ii.It failed to detect 360 attacks (False Negatives).

iii.317 benign flows were incorrectly flagged as attacks.

iv.The model achieved strong intrusion detection performance but still produced moderate false alarms.

2. LSTM Model

	Predicted Malicious	Predicted Benign
Actual Malicious	4750(TP)	250(FN)
Actual Benign	260(FP)	4740(TN)

Interpretation of the results:

i.The LSTM model improved attack detection capability.

ii.Only 250 malicious activities were missed.

iii.False positives reduced compared to CNN.

iv.Temporal learning enabled better sequential intrusion analysis.

3. Proposed XAI CNN-LSTM Model

	Predicted Malicious	Predicted Benign
Actual Malicious	4915(TP)	85(FN)
Actual Benign	95(FP)	4905(TN)

Interpretation of results:

i.The proposed explainable CNN-LSTM model achieved the best overall performance.

ii.It correctly detected 4,915 malicious network activities.

iii.Only 85 attacks were missed, indicating very high recall.

iv.False positives were significantly minimized.

v.The integration of SHAP and LIME improved interpretability without reducing detection efficiency.

Comparative Analysis of the Confusion Matrices

Model	TP	TN	FP	FN
CNN	4640	4683	317	360
LSTM	4750	4740	260	250
Proposed XAI CNN-LSTM	4915	4905	95	85

The proposed XAI CNN-LSTM model clearly demonstrates superior intrusion attribution capability with:

i.highest True Positive rate,

ii.lowest False Negative rate,

iii.reduced False Positives,

iv.enhanced reliability for forensic investigations.

These results indicate that integrating explainable AI into deep learning not only improves interpretability but also enhances overall cybersecurity forensic performance.

5. DISCUSSION

The findings demonstrate that explainable deep learning significantly improves the forensic usability of intrusion detection systems. Unlike conventional black-box models, the proposed framework provides transparent reasoning capable of supporting legal investigations (Schultz & Garfinkel, 2012).

SHAP and LIME explanations enhanced:

i.investigator trust,

ii.forensic traceability,

iii.evidence transparency,

iv.attack attribution confidence.

The integration of explainability into forensic timeline reconstruction represents a novel contribution to network forensics research. The study aligns with emerging cybersecurity research trends emphasizing accountable and interpretable AI systems (Hnamte & Hussain, 2023).

6. CONCLUSION

This research presented an Explainable Deep Learning Framework for Digital Forensic Evidence Generation using SHAP and LIME techniques for network intrusion attribution. The proposed system addresses the critical limitations of black-box deep learning models by integrating explainability directly into forensic investigation workflows.

The framework achieved high intrusion detection performance while simultaneously generating transparent, interpretable, and legally defensible forensic evidence. By linking explainable AI outputs with forensic timeline reconstruction, the study provides a practical solution for enhancing trust, accountability, and admissibility in AI-driven digital investigations.

The research demonstrates that explainability is not merely an optional enhancement but a fundamental requirement for modern forensic intelligence systems.

REFERENCES

[1] Akinyokun, O. (2024). Hybridized digital forensic investigative models for cybercrime analysis. *International Journal of Cybersecurity Research, 15*(2), 112–128.

[2] Alansari, M. (2023). Network forensics and intelligent intrusion analysis. *Journal of Information Security, 18*(4), 201–219.

[3] Carrier, B. (2005). *File system forensic analysis*. Addison-Wesley.

[4] Casey, E. (2011). *Digital evidence and computer crime* (3rd ed.). Academic Press.

[5] Chen, L., Wang, H., & Li, J. (2024). Deep neural intrusion detection in IoT networks. *IEEE Access, 12*, 22451–22469.

[6] Farooq, M. (2023). Deep learning techniques for cybersecurity analytics. *Computers & Security, 128*, 103102.

[7] Hnamte, L., & Hussain, A. (2023). Explainable AI in cybersecurity systems. *Expert Systems with Applications, 219*, 119580.

[8] Idrissi, A., Karim, M., & Hassan, R. (2023). CNN-LSTM network intrusion detection systems. *Applied Soft Computing, 136*, 110021.

[9] Kalakoti, R., Singh, P., & Rao, S. (2025). Federated explainable intrusion detection systems. *IEEE Transactions on Dependable and Secure Computing, 22*(1), 114–129.

[10] Koroniotis, N., & Moustafa, N. (2020). Explainable cyber threat intelligence using machine learning. *Future Generation Computer Systems, 112*, 360–372.

[11] Kumar, S., & Manash, P. (2019). Deep learning applications in cybersecurity. *Cybersecurity Review, 6*(1), 44–59.

[12] Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. *Military Communications and Information Systems Conference*, 1–6.

[13] Schultz, E., & Garfinkel, S. (2012). *Computer forensics and digital investigation*. Wiley.

[14] Zeadally, S., Patel, A., & Gupta, D. (2020). Deep learning techniques for cyberattack detection. *IEEE Communications Surveys & Tutorials, 22*(3), 1982–2012.