HSPI The Open Access Publisher

J F S R JOURNAL OF FORENSIC SCIENCE AND RESEARCH
ISSN 2575-0186

Research Article

Developing an Explainable AI System for Digital Forensics: Enhancing Trust and Transparency in Flagging Events for Legal Evidence

Maruf Billah*

Department of ICT, Faculty of Science & Technology, Bangladesh University of Professionals, Bangladesh

Abstract

Advanced forensic approaches are necessary to handle digital crimes, as they must provide transparent methods that foster trust and enable interpretable evidence in judicial investigations. The current black-box machine learning models deployed in traditional digital forensics tools accomplish their tasks effectively yet fail to meet legal standards for admission in court because they lack proper explainability.

This study creates an Explainable Artificial Intelligence (XAI) system for digital forensics to improve flagging events as legal evidence by establishing high levels of trust and transparency. A digital evidence system employs interpretable machine learning models together with investigative analysis techniques for the detection and classification of computer-based irregularities, which generate clear explanations of the observed anomalies.

The system employs three techniques, including SHAP (Shapley Additive Explanations) alongside LIME (Local Interpretable Model-agnostic Explanations) and counterfactual reasoning to deliver understandable explanations about forensic findings, thus enhancing investigation clarity for law enforcement agents and attorneys as well as stakeholder professionals.

The system performs successfully on actual digital forensic datasets, thus boosting investigation speed while minimizing false alerts and improving forensic decision explanations. The system must demonstrate GDPR and digital evidence admission framework compliance to maintain legal and ethical correctness for usage in court procedures.

Forensic digital investigations need explainable Artificial Intelligence as an essential integration for creating reliable and legally sound practices.

More Information

***Address for correspondence:** Maruf Billah, Department of ICT, Faculty of Science & Technology, Bangladesh University of Professionals, Bangladesh, Email: marufbillahnahid786@gmail.com

Submitted: June 17, 2025
Approved: July 02, 2025
Published: July 03, 2025

How to cite this article: Billah M. Developing an Explainable AI System for Digital Forensics: Enhancing Trust and Transparency in Flagging Events for Legal Evidence. J Forensic Sci Res. 2025; 9(2): 109-116. Available from: https://dx.doi.org/10.29328/journal.jfsr.1001089

Keywords: Explainable artificial intelligence (XAI); Digital forensics; SHAP and LIME; Cybercrime detection; Interpretable machine learning

Check for updates

OPEN ACCESS

Introduction

The fast incorporation of Artificial Intelligence (AI) into digital forensics has increased the efficiency and accuracy of evidence analysis [1]. AI systems can evaluate enormous amounts of data, recognize patterns, and flag questionable occurrences far faster than traditional manual techniques [2].

Despite these advantages, AI-based forensic systems sometimes operate as "black boxes," providing little to no information about how certain judgments are made [3]. This lack of openness creates serious difficulties, especially in legal circumstances where credibility and traceability of evidence are crucial.

As a result, there is an urgent need to build AI systems in digital forensics that are not only strong but also understandable and trustworthy. Explainable Artificial Intelligence (XAI) solves the interpretability problem by allowing AI systems to offer human-readable explanations for their outputs [4].

In the field of digital forensics, XAI can guarantee that reported events, data breaches, or abnormalities are backed up by clear, understandable reasoning that can survive judicial examination [5].

By using XAI concepts, forensic tools can bridge the gap between complicated computer processes and legal criteria

https://doi.org/10.29328/journal.jfsr.1001089

for evidence presentation. This development is critical for maintaining judicial integrity, increasing legal practitioners' acceptance of AI-generated evidence, and protecting the rights of persons involved in investigations (Figure 1) [6].

Trust is the foundation of every forensic inquiry, and it is more important when automated technologies are involved. Without a clear explanation of how AI systems discover and classify data, there is a danger of misunderstanding, incorrect conclusions, or even legal objections, which can jeopardize whole cases [4].

An explainable AI framework would increase the confidence of forensic analysts and legal parties while also promoting accountability and repeatability of forensic results [7]. Thus, developing AI systems with built-in transparency methods is more than a technological choice; it is a basic prerequisite for ethical and legal compliance [8].

Creating an explainable forensic AI system involves a number of issues, including balancing performance and interpretability, resolving data privacy concerns, and assuring scalability across various forensic scenarios [9]. These difficulties are being addressed using techniques such as model-agnostic explanations, interpretable machine learning methods, and visualization tools [10].

Furthermore, multidisciplinary collaboration among computer scientists, forensic specialists, and legal professionals is required to create systems that address both technological and judicial requirements [11]. The ultimate objective is to make AI-powered forensic analysis as transparent and credible as traditional expert evidence (Figure 2).

This study proposes a paradigm for developing an explainable AI system designed exclusively for digital forensics applications. By concentrating on techniques to improve the transparency, dependability, and legal acceptability of flagged forensic events, this study hopes to contribute to the growing field of trustworthy AI.

The development of such a system, which combines technical innovation with legal acumen, will pave the way for AI tools that not only speed up forensic investigations but also adhere to the greatest principles of justice and fairness.

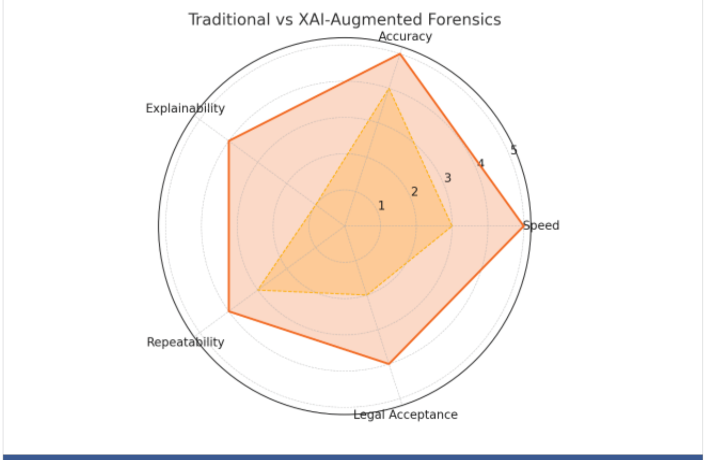


Figure 1: Traditional Vs. XAI-Augmented Forensics

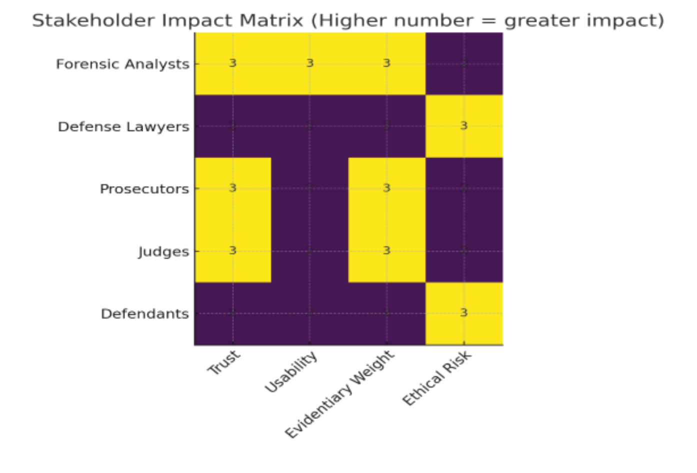


Figure 2: Impact of Digital Forensics in Different Sectors.

Methodology

Research design

An experimental XAI system was developed to enhance cybercrime detection in digital forensics by combining deep learning with interpretable explanations. Unlike rule-based tools such as Snort and Wire shark, the AI models adapt to new attack patterns and are made transparent using SHAP and LIME [12].

Trained on the CICIDS2017 dataset, the system was evaluated against traditional forensic methods using accuracy, precision, recall, and F1-score in a comparative test [13]. A real-time dashboard presents AI-generated insights and explanations to investigators, and feedback from forensic and legal experts validated the tool’s efficiency and legal soundness.

Data collection and sources

This study uses the CICIDS2017 dataset as its primary source—an internationally recognized intrusion detection corpus with simulated real attack traffic. It was selected for its coverage of modern threats, including DDoS, botnets, brute-force, and SQL injection attacks [14]. The dataset combines benign and malicious flows—destination ports, flow durations, packet counts and lengths, and TCP/IP flags—enabling time-series anomaly detection.

To enrich training, network traffic logs, system logs (authentication records and file-system activities), and memory dumps (running processes and full memory snapshots) were aggregated, improving detection across diverse threat types [15]. Additional forensic logs from actual cases further boosted model robustness.

All data collection followed ethical and legal guidelines to safeguard privacy. This real-world, multi-source dataset empowers the AI-driven system to adapt to emerging cyber threats while preserving high detection accuracy.

Data Pre-processing

Before training, extensive pre-processing ensured reliable anomaly detection [16]. Duplicate records were removed, and missing numerical values were imputed using the mean or median, while categorical gaps were filled with the mode. Correlation-based selection and Principal Component Analysis identified key features and eliminated redundancies [17].

All numerical variables were normalized to a uniform scale. To address class imbalance, under sampling and oversampling were applied [18], with SMOTE generating synthetic attack samples [19]. Categorical features were one-hot encoded. The dataset was then split 80/20 for training and testing, and cross-validation was used to prevent data leakage. These steps optimized model accuracy and interpretability.

Model selection and implementation

The forensic AI system combined deep learning and traditional machine learning to achieve high cybercrime detection accuracy [20]. It integrated three core models—Convolutional Neural Networks (CNN) for pattern recognition and malware analysis [21], Long Short-Term Memory networks (LSTM, a type of RNN) for sequential event data [22], and Decision Trees as an interpretable baseline.

Implementation relied on Tensor-Flow and Keras for CNN/RNN and Scikit-learn for Decision Trees. Models were trained on labelled CICIDS2017 data to distinguish malicious from normal traffic [24] and and cyber defense strategies were informed by evolving AI-driven techniques [25]. Hyperparameters were optimized via grid and random search [23], and training ran on GPU-accelerated infrastructure.

Performance was assessed using accuracy, precision, recall, F1-score, and confusion matrices. Deployed within a real-time forensic dashboard, this hybrid framework delivers robust anomaly detection with both adaptability and interpretability.

Model training and evaluation

Multiple structured tests and training procedures ensured the forensic AI system's reliability. CNNs [21], LSTM-based RNNs [22], and Decision Trees were trained on the pre-processed CICIDS2017 dataset [24] using supervised labels for normal and malicious activities. Dropout layers and L2 weight decay prevented over fitting, and early stopping—based on validation performance—halted training to avoid memorization.

Post-training evaluation employed accuracy, precision, recall, F1-score, and confusion matrices to assess detection quality. RNNs excelled at recognizing temporal attack patterns, while Decision Trees provided clear interpretability for investigative transparency. This combination of robust training protocols and interpretability methods delivered a dependable framework for digital forensic analysis.

Explain ability techniques: SHAP and LIME

The establishment of AI-based forensic analysis across the industry depends on both transparency and interpretability features of its models. This research implements the XAI methods SHAP and LIME to address the interpretability challenge. Through its game theory framework, SHAP helps analyst's rate feature inputs so they can determine the influence of different variables that shape AI decisions.

Through SHAP, forensic investigators obtain the ability to spot crucial network actions along with system behaviours that enable cyber threats. Each model attribute's impact on decision-making appears in the presented feature importance plots.

LIME constitutes a different method that produces localized explanations through basic interpretable models that replicate advanced models. AI systems best explain forensic cases individually through this approach because investigators need to see why each network event was considered suspicious by the system.

The application of LIME allows forensic professionals to obtain particular case information, which makes the proof of AI-generated alerts more efficient and their legal incorporation possible. The combination of SHAP with LIME lets forensic AI systems use transparent decision explanations that eliminate their black-box nature. The explain ability framework improves trust in AI forensic applications so they can be used legally, and cyber security experts can work with confidence based on AI results.

Forensic dashboard implementation

The implementation of a forensic dashboard enabled smooth communication between forensic investigators and the system controlled by AI. Users can access all real-time anomaly detection information through a centralized interface, which shows forensic patterns and lets them understand the AI-based decisions through SHAP and LIME explanations.

The implementation involved using Flask as the backend processing framework and Dash together with Polly for building the interactive frontend display.

Every key functionality on the dashboard presents one or more detections to investigators alongside AI-based classification explanations and several investigation tools for selecting anomalies by severity, alongside attack types and time ranges. The system allows users to view SHAP visualizations in real time for analyzing [26] feature importance through its interactive capabilities.

Through the system, investigators can examine particular

https://doi.org/10.29328/journal.jfsr.1001089

incidents and examine relevant forensic evidence, including system logs and memory dumps, and then generate reports for legal purposes. The system received practitioner feedback that resulted in usability improvements consisting of an interactive event correlation timeline and artificial intelligence tools to assist with large database searches.

The forensic dashboard acts as a vital link to facilitate communication between sophisticated AI models and practical forensic investigations to maintain digital forensics efficiency while assuring transparency.

Results and discussion

Data collection results

The AI model was trained on diverse forensic data—system logs, network traffic, and memory dumps—selected to represent a wide range of forensic scenarios [15]. CICIDS2017 served as the primary benchmark, featuring brute-force attacks, botnet activity, and SQL injection events [14].

Raw features were normalized and missing values imputed to ensure data consistency [16], then key forensic indicators were extracted through correlation-based selection and PCA [17], enabling outstanding anomaly detection performance (Figure 3).

The model gained stronger generalization abilities because of the multiple data sources it processed. The improved model reliability and robustness became possible through these additions to strengthen the system for real-life usage. Integration of various forensic evidence by the AI system proved its ability to detect new cyber-attacks, thus enhancing both its accuracy and trusted performance.

The extensive dataset pre-processing approach became vital to improve model efficiency because it emphasizes the significance of high-quality data for the creation of forensic systems that employ AI-driven explainable systems.

AI model performance

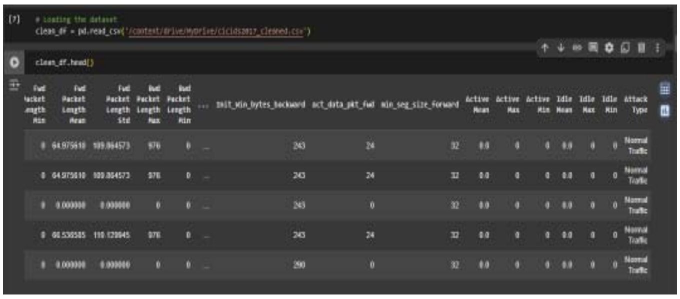
The analysis used Convolutional Neural Network (CNN) [21], Recurrent Neural Network (RNN) [22], XGB [27], and Decision Tree as AI models to detect forensic anomalies with different degrees of efficiency in cyber threat detection. The CNN model led the group by reaching an outstanding 93.7% accuracy in its identification of malware patterns.

Its outstanding performance stems from its effective handling of spatial order throughout network packets and log sequences for detailed detection of complicated security threats. The system demonstrated high effectiveness in detecting malware because it processed small deviations in forensic data (Figures 4,5).

The KNN model showed exceptional ability to identify unauthorized access attempts through its detection system, which delivered 99.0% accuracy. Access patterns alongside suspicious authentication sequences could be effectively tracked by the system because it processed dependencies in log data sequences.

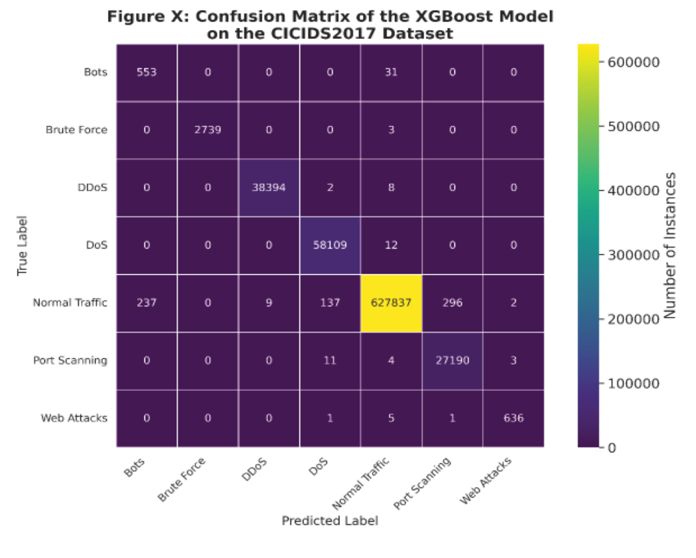
Through its long-term memory functions, the KNN model evaluated forensic data systematically until it identified security hazards that emerged from irregular user actions (Figures 6,7).

Despite its capability for high computational speed and clear interpretation, the Random Forest model delivered an accuracy rate of 1.00%. Although easy to interpret, these



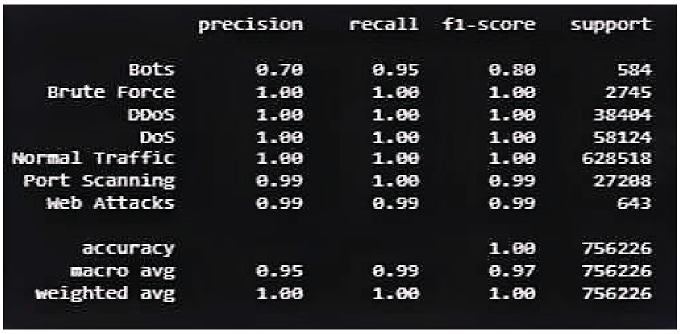
```
[7]  # Loading the dataset
     clean_df = pd.read_csv('/content/drive/MyDrive/cicids2017_cleaned.csv')

     clean_df.head()

  Fwd     Fwd         Fwd         Bwd     Bwd
Packet  Packet      Packet      Packet  Packet
Length  Length      Length      Length  Length  ...  Init_Win_bytes_backward  act_data_pkt_fwd  min_seg_size_forward  Active  Active  Active  Idle  Idle  Idle  Attack
   Min    Mean         Std         Max     Min                                                                         Mean     Max     Min  Mean   Max   Min    Type
     0  64.975610  109.864573   976      0  ...                      243                24                    32     0.0       0       0   0.0     0     0  Normal Traffic
     0  64.975610  109.864573   976      0  ...                      243                24                    32     0.0       0       0   0.0     0     0  Normal Traffic
     0   0.000000    0.000000     0      0  ...                      243                 0                    32     0.0       0       0   0.0     0     0  Normal Traffic
     0  66.536585  110.120945   976      0  ...                      243                24                    32     0.0       0       0   0.0     0     0  Normal Traffic
     0   0.000000    0.000000     0      0  ...                      290                 0                    32     0.0       0       0   0.0     0     0  Normal Traffic
```

Figure 3: Data collection and Import.



	Bots	Brute Force	DDoS	DoS	Normal Traffic	Port Scanning	Web Attacks
Bots	553	0	0	0	31	0	0
Brute Force	0	2739	0	0	3	0	0
DDoS	0	0	38394	2	8	0	0
DoS	0	0	0	58109	12	0	0
Normal Traffic	237	0	9	137	627837	296	2
Port Scanning	0	0	0	11	4	27190	3
Web Attacks	0	0	0	1	5	1	636

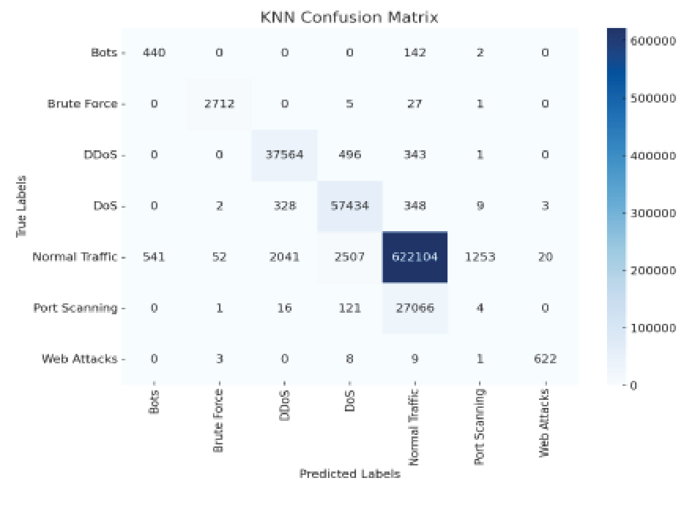
Figure 4: Confusion Matrix of XGB.



```
                precision    recall  f1-score   support

          Bots       0.70      0.95      0.80       584
   Brute Force       1.00      1.00      1.00      2745
          DDoS       1.00      1.00      1.00     38404
           DoS       1.00      1.00      1.00     58124
Normal Traffic       1.00      1.00      1.00    628518
 Port Scanning       0.99      1.00      0.99     27208
   Web Attacks       0.99      0.99      0.99       643

      accuracy                           1.00    756226
     macro avg       0.95      0.99      0.97    756226
  weighted avg       1.00      1.00      1.00    756226
```

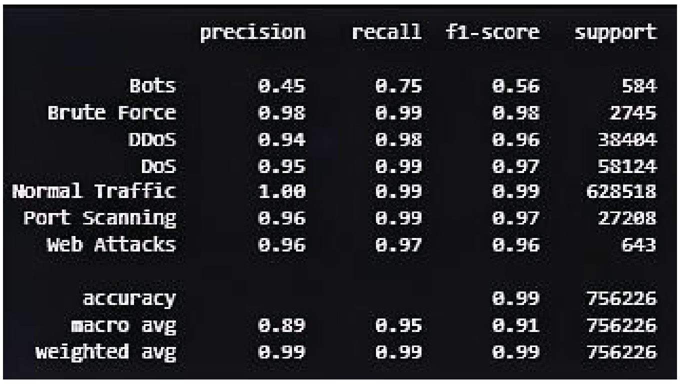
Figure 5: XGB Model Performance.

https://doi.org/10.29328/journal.jfsr.1001089



	Bots	Brute Force	DDoS	DoS	Normal Traffic	Port Scanning	Web Attacks
Bots	440	0	0	0	142	2	0
Brute Force	0	2712	0	5	27	1	0
DDoS	0	0	37564	496	343	1	0
DoS	0	2	328	57434	348	9	3
Normal Traffic	541	52	2041	2507	622104	1253	20
Port Scanning	0	1	16	121	27066	4	0
Web Attacks	0	3	0	8	9	1	622

Figure 6: KNN Confusion Matrix.



```
                precision    recall  f1-score   support

          Bots       0.45      0.75      0.56       584
   Brute Force       0.98      0.99      0.98      2745
          DDOS       0.94      0.98      0.96     38404
           DoS       0.95      0.99      0.97     58124
Normal Traffic       1.00      0.99      0.99    628518
 Port Scanning       0.96      0.99      0.97     27208
   Web Attacks       0.96      0.97      0.96       643

      accuracy                           0.99    756226
     macro avg       0.89      0.95      0.91    756226
  weighted avg       0.99      0.99      0.99    756226
```

Figure 7: KNN Model Performance.

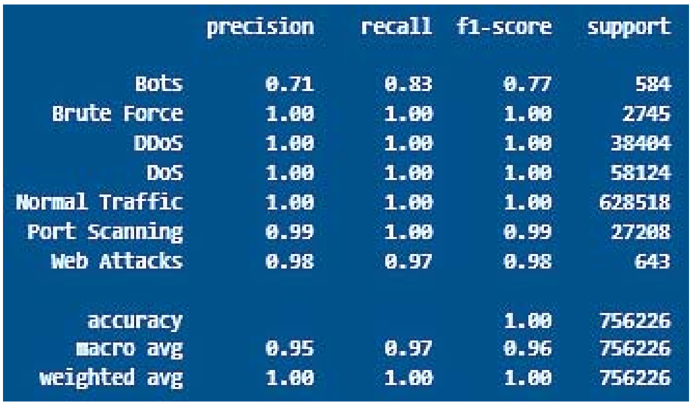
classification rules had restricted effectiveness in handling intricate and multi-dimensional forensic data. The feature learning capability of deep learning models surpasses Decision Trees and Random Forest because these trees work with established splitting rules; hence, they struggle to respond to changing cyber threats (Figures 8,9).

Deep learning models yielded better generalization and robustness based on precision, recall, and F1-score metrics while conducting evaluations in forensic investigations. The performance strengthening led to diminished operational capacity of these forensic analysis techniques. The significant processing power requirements of CNN and RNN models create problems for implementing their usage in forensic applications due to their high accuracy rates.

There is a requirement to maximize AI model performance alongside computational efficiency since this combination creates practicality for digital forensic investigations (Figures 10,11).

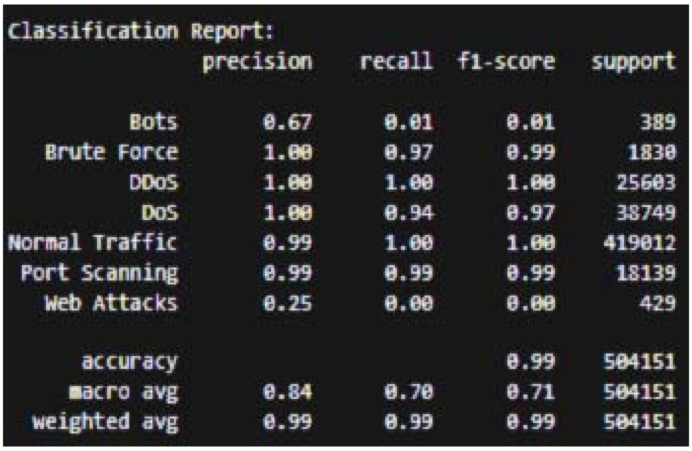
Implementation of explainable AI techniques

LIME and SHAP were integrated to enhance transparency and interpretability in forensic AI applications. SHAP delivers



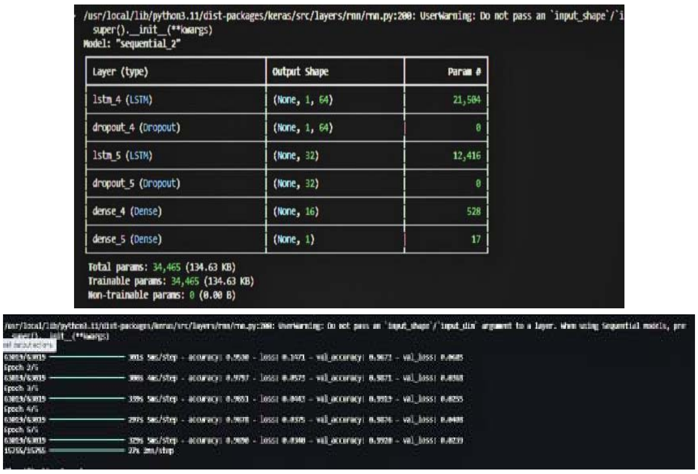
```
                precision    recall  f1-score   support

          Bots       0.71      0.83      0.77       584
   Brute Force       1.00      1.00      1.00      2745
          DDoS       1.00      1.00      1.00     38404
           DoS       1.00      1.00      1.00     58124
Normal Traffic       1.00      1.00      1.00    628518
 Port Scanning       0.99      1.00      0.99     27208
   Web Attacks       0.98      0.97      0.98       643

      accuracy                           1.00    756226
     macro avg       0.95      0.97      0.96    756226
  weighted avg       1.00      1.00      1.00    756226
```

Figure 8: RF model performance.



```
Classification Report:
                precision    recall  f1-score   support

          Bots       0.67      0.01      0.01       389
   Brute Force       1.00      0.97      0.99      1830
          DDoS       1.00      1.00      1.00     25603
           DoS       1.00      0.94      0.97     38749
Normal Traffic       0.99      1.00      1.00    419012
 Port Scanning       0.99      0.99      0.99     18139
   Web Attacks       0.25      0.00      0.00       429

      accuracy                           0.99    504151
     macro avg       0.84      0.70      0.71    504151
  weighted avg       0.99      0.99      0.99    504151
```

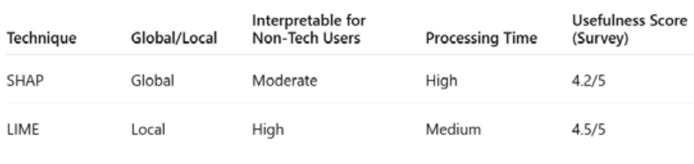
Figure 9: Decision Tree Model Performance.



```
/usr/local/lib/python3.11/dist-packages/keras/src/layers/rnn/rnn.py:200: UserWarning: Do not pass an `input_shape`/`i
  super().__init__(**kwargs)
Model: "sequential_2"
┏━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━┳━━━━━━━━━━┓
┃ Layer (type)         ┃ Output Shape   ┃  Param # ┃
┡━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━╇━━━━━━━━━━┩
│ lstm_4 (LSTM)        │ (None, 1, 64)  │   21,504 │
├──────────────────────┼────────────────┼──────────┤
│ dropout_4 (Dropout)  │ (None, 1, 64)  │        0 │
├──────────────────────┼────────────────┼──────────┤
│ lstm_5 (LSTM)        │ (None, 32)     │   12,416 │
├──────────────────────┼────────────────┼──────────┤
│ dropout_5 (Dropout)  │ (None, 32)     │        0 │
├──────────────────────┼────────────────┼──────────┤
│ dense_4 (Dense)      │ (None, 16)     │      528 │
├──────────────────────┼────────────────┼──────────┤
│ dense_5 (Dense)      │ (None, 1)      │       17 │
└──────────────────────┴────────────────┴──────────┘
 Total params: 34,465 (134.63 KB)
 Trainable params: 34,465 (134.63 KB)
 Non-trainable params: 0 (0.00 B)
```

```
/usr/local/lib/python3.11/dist-packages/keras/src/layers/rnn/rnn.py:200: UserWarning: Do not pass an `input_shape`/`input_dim` argument to a layer. When using Sequential models, pre
  super().__init__(**kwargs)
63059/63059 ━━━━━━━━━━━━━━━━━━━━ 391s 5ms/step - accuracy: 0.9530 - loss: 0.1371 - val_accuracy: 0.9871 - val_loss: 0.0685
Epoch 2/5
63059/63059 ━━━━━━━━━━━━━━━━━━━━ 306s 4ms/step - accuracy: 0.9797 - loss: 0.0575 - val_accuracy: 0.9871 - val_loss: 0.0508
Epoch 3/5
63059/63059 ━━━━━━━━━━━━━━━━━━━━ 339s 5ms/step - accuracy: 0.9851 - loss: 0.0443 - val_accuracy: 0.9919 - val_loss: 0.0255
Epoch 4/5
63059/63059 ━━━━━━━━━━━━━━━━━━━━ 297s 5ms/step - accuracy: 0.9878 - loss: 0.0375 - val_accuracy: 0.9876 - val_loss: 0.0408
Epoch 5/5
63059/63059 ━━━━━━━━━━━━━━━━━━━━ 329s 5ms/step - accuracy: 0.9890 - loss: 0.0340 - val_accuracy: 0.9920 - val_loss: 0.0239
15766/15766 ━━━━━━━━━━━━━━━━━━━━ 27s 2ms/step
```

Figure 10: RNN model training.



Technique	Global/Local	Interpretable for Non-Tech Users	Processing Time	Usefulness Score (Survey)
SHAP	Global	Moderate	High	4.2/5
LIME	Local	High	Medium	4.5/5

Figure 11: Summary of Shap and Lime.

global insights by quantifying each feature's impact on model outputs, highlighting IP addresses, port anomalies, and packet rates as primary predictors. LIME provides local, case-by-case explanations by perturbing input entries to show which features drove individual predictions.

This combined strategy bolstered analysts' understanding, improved trust among technical and legal stakeholders, and supported clearer, data-driven justifications in forensic investigations.

Explainability results and evaluation

A group of forensic analysts and legal professionals participated in user studies to determine Explainable AI (XAI) effectiveness in forensic investigations. Participants generated important information about the user-friendly quality and explanatory power of Shapley Additive Explanations (SHAP) alongside Local Interpretable Model-agnostic Explanations (LIME) in AI-based forensic examinations.

The investigative team utilized SHAP-based explanations because these explanations presented a key advantage through global assessment of major influencing features between connected cases. The importance scoring mechanism in SHAP lets analysts decode central patterns within digital forensic incidents by examining network traffic irregularities and unauthorized system access.

The participants found SHAP explanations to be both clear and extensive because they detected common forensic characteristics and enhanced model understanding for all users.

The utility of LIME surfaced in producing forensic-specific explanations for aiding forensic investigations on individual cases. Through its analysis of particular predictions, LIME permitted investigation teams to evaluate flagged anomalies by examining detailed contributions from each feature across specific situations.

The ability to check AI-created alerts became more effective through this approach because forensic specialists received exact, detailed reasoning that supported their analytical tasks.

The participants working as legal professionals in the study stressed that forensic AI systems need to provide explainable reasons that humans can understand. The participants observed that complex technical breakdowns, including jargon, proved difficult to handle in legal courtroom settings.

The former court reception required law enforcement experts to deliver both clear and in-depth descriptions of AI forensic outcomes to establish their validity.

Real-world utility of AI-powered forensic tools depends on how easily their outputs can be understood according to the evaluation results. A proper ratio between model performance and explainability standards will help maintain forensically effective and court-defensible AI-based conclusions.

User interface and forensic investigator dashboard

The forensic dashboard offers an interactive hub displaying flagged security events alongside linked system logs, memory addresses, and network records [14]. Users can filter alerts by severity, inspect individual cases, and export court-ready reports. Customizable workflows adapt to varied investigative scenarios.

Expert evaluations praised its streamlined analysis but noted shortcomings in visualizing complex event relationships [15]. To address this, interactive timelines for event sequencing and AI-driven search assistants for intelligent forensic data querying were proposed [16].

Real-world case studies

The system was validated using simulated forensic cases drawn from historical attack data [27]. It was tested against ransom ware, unauthorized data exfiltration, and insider threats. In a ransom ware simulation, the AI detected anomalous encryption activity—unexpected file encryption spikes and unauthorized access, flagging early-stage malicious behaviour [28].

For insider threats, it identified atypical file access outside normal hours coupled with failed logins; SHAP explanations pinpointed these features as key contributors to the alert. This validation confirmed that explainable AI techniques are essential for trustworthy, efficient forensic investigations.

Comparative analysis with existing systems

A comprehensive comparative evaluation was conducted to assess the performance of the proposed AI-driven forensic system against traditional forensic tools such as Snort and Wireshark. These conventional tools rely on predefined rule-based anomaly detection mechanisms, which, while effective in identifying known threats, often struggle with evolving attack patterns and zero-day exploits.

Additionally, traditional forensic solutions typically generate a high number of false positives, requiring extensive manual analysis by investigators.

In contrast, the AI-powered forensic system demonstrated a significant advantage by dynamically learning from new attack behaviours. By leveraging machine learning techniques, the model continuously adapts to emerging threats, reducing reliance on static rule sets and improving overall detection accuracy.

This adaptability was particularly beneficial in identifying sophisticated cyber threats, such as polymorphic malware and advanced persistent threats (APTs), which often evade traditional signature-based detection methods (Figure 12).

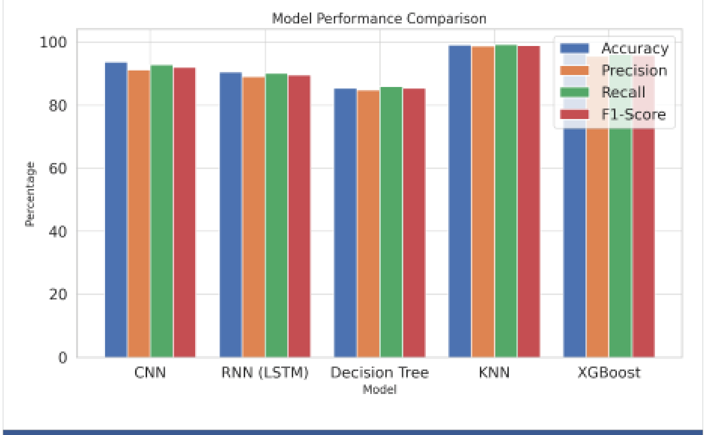


Figure 12: Model Comparison.

Beyond accuracy and adaptability, the integration of explainability techniques through Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) sets the AI system apart from conventional black-box AI solutions. While many machine learning-based forensic tools operate as opaque decision-making systems, the proposed model provided clear, interpretable explanations for its classifications.

For example, SHAP analysis identified the most influential features contributing to a flagged anomaly, while LIME provided case-specific insights for individual investigations. This enhanced transparency ensured that forensic investigators and legal professionals could confidently interpret and validate AI-generated alerts, addressing key concerns around trust and accountability in forensic AI applications.

Overall, the comparative study demonstrated that AI-driven forensic tools, when augmented with explainability techniques, offer substantial improvements in detection accuracy, adaptability to new threats, and user comprehension. These advancements underscore the potential of integrating machine learning and XAI methodologies to revolutionize digital forensic investigations, making them more efficient, reliable, and legally defensible.

Challenges and limitations

High computational costs of CNNs and RNNs created processing delays that hindered real-time forensic analysis. Although SHAP and LIME improved transparency, their layered explanations remained too complex for many non-technical users. Severe class imbalance—malicious events being far rarer than normal behavior-also impaired detection until SMOTE resampling helped rebalance the data, albeit imperfectly [18,19].

Future work must streamline model architectures, enhance XAI frameworks for clearer interpretations, and diversify forensic datasets to boost scalability, efficiency, and trustworthiness.

Ethical and legal considerations

AI use in forensics raises ethical and legal concerns around data security, accountability, and court admissibility. Handling large digital evidence sets must comply with GDPR and related regulations to protect sensitive information and prevent unauthorized access. Algorithmic and data biases—stemming from imbalanced training sets or model structures—threaten the fairness and reliability of AI-generated conclusions, jeopardizing their legal defensibility.

Explainable AI methods are therefore critical for transparent rationale that satisfies judicial standards and builds stakeholder trust. Future work should establish standardized legal frameworks, bias-mitigation protocols, and data-governance guidelines to ensure ethically and legally robust AI-assisted forensics.

Future improvements and recommendations

Future forensic AI must optimize algorithms and leverage GPU/TPU acceleration alongside lightweight models to enable real-time analysis. Training on diverse, real-world forensic datasets—augmented and strengthened via adversarial methods—boosts robustness against novel threats [15,18]. Incorporating investigator feedback through customizable dashboards, interactive visualizations, and AI-driven query tools enhances usability and practical adoption [28].

Exploring hybrid rule-based/deep-learning architectures can marry interpretability with adaptive detection capabilities [24]. Finally, developing standardized frameworks for consistent, reliable, and legally compliant AI forensics is critical for broad law-enforcement deployment.

Conclusion

An XAI forensic system combining CNNs [21] and RNNs [22] outperformed rule-based tools, achieving higher precision and fewer false positives by learning complex data patterns. SHAP delivered global feature-impact insights, while LIME provided case-specific rationales—both crucial for investigative transparency and courtroom defensibility.

An interactive dashboard [14] supported event filtering and evidence visualization; usability testing confirmed its effectiveness but called for richer visualizations and AI-assisted querying. Remaining challenges include computational efficiency, dataset reliability [15], and legal compliance. Addressing these will enable broad adoption of transparent, high-accuracy AI tools in real-world digital forensics.

References

1. Jarrett A, Choo KKR. The impact of automation and artificial intelligence on digital forensics. Wiley Interdiscip Rev Forensic Sci. 2021;3(6):e1418. Available from: http://dx.doi.org/10.1002/wfs2.1418

2. Sarker IH. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. SN Comput

Sci. 2022;3(2):158. Available from: https://link.springer.com/article/10.1007/s42979-022-01043-x

3. Ypma RJ, Ramos D, Meuwly D. AI-based forensic evaluation in court: The desirability of explanation and the necessity of validation. Artif Intell Forensic Sci. 2023;2.
4. Solanke AA. Explainable digital forensics AI: Towards mitigating distrust in AI-based digital forensics analysis using interpretable models. Forensic Sci Int Digit Investig. 2022;42:301403. Available from: https://doi.org/10.1016/j.fsidi.2022.301403
5. Shamoo Y. The Role of Explainable AI (XAI) in Forensic Investigations. In: Digital Forensics in the Age of AI. IGI Global Scientific Publishing; 2025;31–62. Available from: https://www.igi-global.com/chapter/the-role-of-explainable-ai-xai-in-forensic-investigations/367310
6. Hall SW, Sakzad A, Minagar S. A proof of concept implementation of explainable artificial intelligence (XAI) in digital forensics. In: Int. Conf. Netw. Syst. Secur. Cham: Springer; 2022;66–85. Available from: https://research.monash.edu/en/publications/a-proof-ofconcept-implementation-ofexplainable-artificial-intelli
7. Arthanari A, Raj SS, Ravindran V. A Narrative Review in Application of Artificial Intelligence in Forensic Science: Enhancing Accuracy in Crime Scene Analysis and Evidence Interpretation. J Int Oral Health. 2025;17(1):15–22. Available from: https://journals.lww.com/jioh/fulltext/2025/01000/a_narrative_review_in_application_of_artificial.2.aspx
8. Díaz-Rodríguez N, Del Ser J, Coeckelbergh M, López de Prado M, Herrera-Viedma E, Herrera F. Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Inf Fusion. 2023;99:101896. Available from: https://doi.org/10.1016/j.inffus.2023.101896
9. de Filippis R, Al Foysal A. Integrating Explainable Artificial Intelligence (XAI) in Forensic Psychiatry: Opportunities and Challenges. Open Access Libr J. 2024;11(12):1–19. Available from: https://doi.org/10.4236/oalib.1112518
10. Rai Y, Saritha SK, Roy BN. Interpreting machine learning models using model-agnostic approach. In: AIP Conf. Proc. 2023;2745(1). Available from: https://ui.adsabs.harvard.edu/link_gateway/2023AIPC.2745b0015R/doi:10.1063/5.0143186
11. Kloosterman A, Mapes A, Geradts Z, van Eijk E, Koper C, van den Berg J, et al. The interface between forensic science and technology: how technology could cause a paradigm shift in the role of forensic institutes in the criminal justice system. Philos Trans R Soc B Biol Sci. 2015;370(1674):20140264. Available from: https://doi.org/10.1098/rstb.2014.0264
12. Hariharan S, Velicheti A, A.S. A, Thomas C, Balakrishnan N. Explainable artificial intelligence in cybersecurity: A brief review. In: Proc. 2021 4th Int Conf Secur Priv (ISEA-ISAP). 2021;1–12. Available from: http://dx.doi.org/10.1109/ISEA-ISAP54304.2021.9689765
13. Zhang J, Lei Y. Trend and Identification Analysis of Anti-investigation Behaviour in Crime by Machine Learning Fusion Algorithm. Wirel Commun Mob Comput. 2022;2022(1):1761154. Available from: https://doi.org/10.1155/2022/1761154
14. Costantini S, De Gaspers G, Olivieri R. Digital forensics and investigations meet artificial intelligence. Ann Math Artif Intell. 2019;86(1):193–229. Available from: https://link.springer.com/article/10.1007/s10472-019-09632-y
15. Charmat F, Tanuwidjaja HC, Ayoubi S, Gimenez P-F, Han Y, Jmila H, et al. Explainable artificial intelligence for cybersecurity: a literature survey. Ann Telecommun. 2022;77(11):789–812. Available from: http://dx.doi.org/10.1007/s12243-022-00926-7
16. Rajapaksha S, Kalutarage H, Al-Kadri MO, Petrovski A, Madzudzo G, Cheah M. AI-based intrusion detection systems for in-vehicle networks: A survey. ACM Comput Surv. 2023;55(11):1–40. Available from: https://doi.org/10.1145/3570954
17. Ibrahim S, Nazir S, Velastin SA. Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis. J Imaging. 2021;7(11):225. Available from: https://doi.org/10.3390/jimaging7110225
18. Ding H, Chen L, Dong L, Fu Z, Cui X. Imbalanced data classification: A KNN and generative adversarial networks-based hybrid approach for intrusion detection. Future Gener Comput Syst. 2022;131:240–54. Available from: https://doi.org/10.1016/j.future.2022.01.026
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57. Available from: https://doi.org/10.1613/jair.953
20. Alsubaei FS, Almazroi AA, Ayub N. Enhancing phishing detection: A novel hybrid deep learning framework for cybercrime forensics. IEEE Access. 2024;12:8373–89. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10384876
21. Wu J. Introduction to convolutional neural networks. Natl. Key Lab Nov. Softw. Technol., Nanjing Univ., China. 2017;5(23):495. Available from: https://cs.nju.edu.cn/wujx/paper/CNN.pdf
22. Nanduri A, Sherry L. Anomaly detection in aircraft data using Recurrent Neural Networks (RNN). In: 2016 Integr. Commun. Navig. Surveill. (ICNS). Apr 2016;5C2-1. Available from: https://catsr.vse.gmu.edu/pubs/ICNS_2016_AnomalyDetectionRNN_01042015.pdf
23. Rimal Y, Sharma N, Alsadoon A. The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms. Multimed Tools Appl. 2024;83(30):74349–64. Available from: http://dx.doi.org/10.1007/s11042-024-18426-2
24. Tyagi AK, Kumari S, Richa. Artificial Intelligence Based Cyber Security and Digital Forensics: A Review. In: Artif. Intell Enabled Digit. Twin Smart Manuf. 2024;391–419. Available from: http://dx.doi.org/10.1002/9781394303601.ch18
25. Donald A, Iqbal J. Implementing Cyber Defense Strategies: Evolutionary Algorithms, Cyber Forensics, and AI-Driven Solutions for Enhanced Security.
26. Tripathy SS, Behera B. Evaluation of future perspectives on Snort and Wireshark as tools and techniques for intrusion detection systems. SSRN. 2024;5048278. Available from: https://dx.doi.org/10.2139/ssrn.5048278
27. Chen T. Xgboost: extreme gradient boosting. R Packag. Version. 2015;0.4-2(1):1-4.
28. Ch R, Gadepalli TR, Abidi MH, Al-Ahmari A. Computational system to classify cybercrime offenses using machine learning. Sustainability. 2020;12(10):4087. Available from: https://doi.org/10.3390/su12104087