



International Journal of Advanced Research in Computational Thinking and Data Science

Journal homepage:
<https://karyailham.com.my/index.php/ctds/index>
ISSN: 3030-5225



Design and Development Aethertrace: An AI-Powered Chat-Log Analysis for Digital Forensics Investigations

Thivya Kumari Thorais¹, Nur Fasihah Mohd Esa^{1,*}, Asmaa Mahfoud Hezam Al-Hakimi¹

¹ Faculty of Information Sciences and Engineering at Management and Science University, Shah Alam Selangor, Malaysia

ARTICLE INFO

Article history:

Received 9 June 2026

Received in revised form 12 August 2026

Accepted 1 September 2026

Available online 5 October 2026

Keywords:

Digital forensics; chat-log analysis; artificial intelligence; system design and development; AetherTrace

ABSTRACT

The rapid expansion of digital communication platforms has created a massive volume of chat-log evidence that exceeds the capacity for manual forensic review. This research evaluates the efficacy of an automated framework designed to improve the efficiency of identifying suspicious communication patterns in large-scale chat datasets. The system is grounded in the Digital Forensic Process Model, which integrates machine learning directly into the investigation and analysis phases to automate evidence discovery. The study employed a quantitative experimental design using the Plagiarism Analysis, Author Identification, and Near-Duplicate Detection 2012 dataset consisting of 66,000 conversation documents. Data analysis involved a comparative evaluation of Logistic Regression, Multi-Layer Perceptron, and Bidirectional Long Short-Term Memory architectures against a 33 per cent testing split. Bidirectional Long Short-Term Memory and Logistic Regression models both achieved predictive performance for predatory behavior with F1-scores of 0.70. While all models maintained overall accuracy above 0.98, the XGBoost model exhibited significantly lower performance in identifying specific predatory cues with an F1-score of only 0.57. Conversely, XGBoost and Logistic Regression demonstrated the greatest precision in avoiding misclassification with a false positive rate of only 0.2 per cent. This research is original for its novel integration of model interpretability tools to decompose specific linguistic cues during automated forensic triage. These findings provide forensic practitioners with a scalable method to reduce investigation backlogs while maintaining the transparency required for judicial admissibility.

1. Introduction

Global digitalisation has triggered an explosion of digital crimes, with the annual global cost of cybercrime reaching 8.15 trillion USD in 2023. This surge has transformed the investigative landscape, forcing law enforcement to process massive volumes of textual data and metadata from digital communication platforms [1]. Chat logs are now a primary evidence source for serious offences, including child exploitation, where UK police alone received 1,600 reports of online child predatory behaviour within a single year. However, the sheer volume and variety of this evidence often exceed human processing capacity. The reliance on manual forensic review creates significant investigative

* Corresponding author.

E-mail address: fasihah@msu.edu.my

backlogs, which can lead to miscarriages of justice or the discontinuation of cases due to evidence disclosure failings. Consequently, there is an urgent societal need for scalable solutions that can automate evidence discovery without sacrificing the forensic integrity required for legal admissibility [2].

Current research has established that machine learning (ML) and artificial intelligence (AI) can fundamentally enhance digital forensics by identifying anomalies, detecting malware, and automating pattern recognition. Specifically, Natural Language Processing (NLP) has been successfully applied to detect suspicious communication patterns and classify messages as predatory or non-predatory. Integrated frameworks like the Digital Forensic Process Model (DFPM) provide a structured lifecycle encompassing identification, collection, and analysis to ensure that evidence remains admissible in court. Furthermore, advancements in Optical Character Recognition (OCR) allow for the extraction of textual insights from image-based evidence, facilitating a more comprehensive view of digital interactions. Existing studies also indicate that deep learning architectures, such as neural networks, can achieve high accuracy in reconstructing timelines and ranking the relevance of file artefacts, thereby reducing the psychological burden on human investigators who may be exposed to distressing evidence [3].

Despite these advancements, a significant misalignment persists between abstract forensic models and their concrete implementation in real-world investigative workflows. Most contemporary forensic tools remain focused on evidence acquisition and data extraction rather than the intelligent analysis of communication content. This focus creates a theoretical puzzle where investigators possess the necessary data but lack an automated mechanism to interpret the "next step" within a specific case context. While machine learning models can classify data, there is a distinct absence of methodologies that explicitly integrate these models into the deterministic phases of forensic triage to provide explainable results for judicial review. The "black box" nature of some deep learning architectures further complicates this issue, as forensic experts must be able to decompose a model's decision-making process to satisfy the transparency requirements of Art. 6 ECHR. The mechanism for balancing automated speed with the granular interpretability required for court testimony remains unresolved [4].

This study presents AetherTrace, an AI-powered digital forensic system designed to automate the analysis of chat logs and media evidence. Grounded in the Digital Forensic Process Model (DFPM), the system's purpose is to facilitate the automatic discovery of harmful behaviours while maintaining a documented chain of evidence. The research follows an Agile software development design, allowing for the iterative integration of Python-based NLP threat detection and Tesseract-based OCR technologies. The sample consists of the PAN-12 dataset, a standardised corpus of 66,000 documents specifically designed for online predator identification [5]. The system architecture employs an ElectronJS frontend and a FastAPI backend, utilizing SQLite for the centralised storage of investigation records and evidence metadata. This setting serves as a theoretically informative test site because the PAN-12 data is highly imbalanced, mimicking real-world conditions where predatory chats represent a small percentage of a massive volume of consensual communications [6]. Our analytical approach involves integrating NoraAI for forensic guidance and SHA256 hashing to support chain-of-custody verification [7].

This study makes three primary contributions to the field of cybercriminalistics. First, it extends existing literature by integrating automated ML processes directly into the investigative tasks of the DFPM, providing a replicable workflow for large-scale textual triage. Second, it refines the understanding of online predator detection by demonstrating how model interpretability tools can be used to explain automated findings to non-technical judicial stakeholders [8]. Finally, this research validates the use of centralised AI-assisted platforms in reducing investigation backlogs while

ensuring the forensic integrity of chat-log evidence through automated reporting and hashing [9]. The manuscript is organised as follows. Section 2 details the literature review; Section 3 presents the methodology and system architecture; Section 4 presents the experimental results and model performance; and Section 5 discusses the practical implications for forensic practitioners before concluding with future research directions in Section 6.

2. Literature Review

Current scholarship increasingly focuses on the integration of autonomous analytical layers within digital forensic workflows to address the 8.15 trillion USD global cost of cybercrime. This section synthesises empirical evidence regarding the application of machine learning (ML) and natural language processing (NLP) to suspicious communication detection, structured according to the evidence matrix below.

2.1 Evidence Matrix of Empirical Studies

Table 1
Evidence matrix of empirical studies

Authors and Year	Journal and Quartile	Country and Industry	Theory Used	Design and Time Points	N	Measurement Scale/Source	Reported Effect (Statistic)	Unresolved Mechanism	DOI
Ngejane <i>et al.</i> , (2021)	FSI: Digital Investigation (Q1)	South Africa / Law Enforcement	DFPM	Quantitative Experimental	66,000	TF-IDF n-grams (PAN-12)	LR and BiLSTM F1 = 0.70	Cross-platform/language performance	10.1016/j.fsidi.2021.301109
Vyas <i>et al.</i> , (2025)	Procedia Computer Science (Q2)	Global / Social Media	Supervised Learning	Experimental / 3 subtasks	N/A	OLID Dataset	RF/DT Accuracy = 99%	Textual feature impact on accuracy	10.1016/j.procs.2024.11.229
Abdallah <i>et al.</i> , (2025)	Procedia Computer Science (Q2)	Jordan / Digital Forensics	Text Mining	Survey / SLR	66,000	PAN-12 Accuracy Benchmark	NLP Accuracy = 98%	Ethical bias in offender identification	10.1016/j.procs.2024.04.126
Hameed <i>et al.</i> , (2025)	J. Net. Comp. Applications (Q1)	Australia / Cloud Forensics	AZOF Framework	Empirical / AWS Simulated	N/A	Reliability Success Rate	Reliability = 99.8%	Latency in resource-heavy ZKP	10.1016/j.jnca.2024.104331
Sarkar and Shukla (2023)	J. Economic Criminology (Q2)	India / Cybersecurity	Grounded Theory	SLR / 5 stages	237 (Screened)	Open/Axial Coding	N/A	Standardization of SOPs	10.1016/j.jeconc.2023.100034
Chiong <i>et al.</i> , (2021)	Comp. Bio. & Med (Q1)	Malaysia / Health/Forensics	Mixed Models	Experimental / Twitter data	N/A	Depression Textual Features	Successful detection n w/o keywords	Generalizability across unrelated datasets	10.1016/j.compbiomed.2021.104499
Fernandes and Antunes (2023)	FSI: Digital Investigation (Q1)	Portugal / Multimedia	Benford's Law	Statistical Correlation	18,000	Pearson/Spearman Distributions	Average F1 = 90.40%	Portability and GPU card expense	10.1016/j.fsidi.2023.301515
Almuqren <i>et al.</i> , (2024)	Procedia Computer Science (Q2)	Saudi Arabia / Mobile DF	SLR	Systematic Review / PRISMA Benchmarking / AutoDFBench	20 (Studies)	F1-Score Metrics	Malware F1 = 91.34%	Replicability of Android OS images	10.1016/j.procs.2024.01.132
Wickramasekara <i>et al.</i> , (2025)	FSI: Digital Investigation (Q1)	Ireland / AI Forensics	LLM-as-Assistant		N/A	NIST CFTT Datasets	Drone anomaly Accuracy = 92.5%	Measuring investigation verification	10.1016/j.fsidi.2024.301859

Cho <i>et al.</i> , (2025)	FSI: Digital Investigation (Q1)	South Korea / AI Forensics	Forensic Artifact Analysis	Case Study / Framework	N/A	ChatGPT/Claude Extraction Tool	N/A	Forensic security for real-time search	10.1016/j.fsidi.2024.301855
Heitmann and Johnsen (2026)	FSI: Digital Investigation (Q1)	Norway / Law Enforcement	Cognitive Support	Qualitative Case Management	10	Case Management Hierarchy	N/A	Interoperability of TLP frameworks	10.1016/j.fsidi.2026.302111
Shafee <i>et al.</i> , (2025)	Expert Syst. Appl. (Q1)	Portugal / Cybersecurity	OSINT Text Class	Experimental Evaluation	N/A	Stanford Alpaca/Falcon	High classification/Low NER	specialised model entity recognition	10.1016/j.eswa.2024.125509
Dunsin <i>et al.</i> , (2024)	FSI: Digital Investigation (Q1)	UK / Incident Response	AI-DFIR Lifecycle	Systematic Review	N/A	AI/ML Integration Maturity	68% Increase in Cyberattacks	Ethical/Moral privacy Reconstruction	10.1016/j.fsidi.2023.301675
Soni <i>et al.</i> , (2024)	FSI: Digital Investigation (Q1)	India / Remote Access	RAA Forensic Analysis	Multi-platform Analysis	N/A	AnyDesk/Team Viewer Logs	N/A	Cross-source correlation strategy	10.1016/j.fsidi.2024.301838
Laksmi <i>et al.</i> , (2025)	J. Net. Comp. Applications (Q1)	Global / DF Subdomains	theoretical metamodel	Systematic Review / SLR	N/A	Subdomain ambiguous mapping	67% World Internet Usage	Standardized Jurisdictional cooperation	10.1016/j.jnca.2025.104296
Silva <i>et al.</i> , (2025)	FSI: Digital Investigation (Q1)	Brazil / IoT Forensics	SLR / PRISMA	Systematic Review	23 (Final set)	StArt tool quality index	52.17% evaluated models	Multi-protocol IoT adaptability	10.1016/j.fsidi.2024.301912
Stoykova (2024)	Computer Law & Security (Q1)	Norway / Legal DF	RPA Governance	Qualitative Socio-legal	N/A	Reliability Crisis Analysis	1,600 grooming reports/year	ex ante IT validation systems	10.1016/j.clsr.2024.106040
Atefi <i>et al.</i> , (2023)	AAAI Conference (N/A)	Global / Decision Support	Markov Decision	Monte Carlo Tree Search	N/A	K-Nearest Neighbor Reg.	N/A	Principled data-driven support	N/A
Manolache <i>et al.</i> , (2022)	[Source history]	Global / Dark Web	Authorship Verification	Benchmark Dataset	N/A	VeriDark Linguistic Features	N/A	Literary vs Dark Web linguistic delta	N/A
Tallón-Ballesteros (2014)	[Source history]	Spain / Data Mining	Cross-validation	Stratified 4-fold	N/A	C4.5 Decision Trees	Accuracy = 68%	Stochastic outcomes vs Single Output	N/A
Khapse <i>et al.</i> , (2026)	Neurocomputing (Q1)	Global / Agentic AI	Multi-agent Arch.	Systematic Review / PRISMA	237 (Final set)	OAS/BIE complementarity aspects	150+ LLM Safety papers	Human-in-the-loop efficiency tension	10.1016/j.neucom.2026.134049
Alshabib <i>et al.</i> , (2022)	J. Theor. Appl. Inf. Tech. (Q2)	Saudi Arabia / Social Media	DNN Approach	Experimental	N/A	Forensic Analytics Dashboards	N/A	DNN Generalizability	N/A
Karie <i>et al.</i> , (2019)	[Source history]	Global / DF Reporting	Standardization Theory	Descriptive Analysis	N/A	Report Generation Metrics	N/A	Standardizing forensic reports	10.1016/j.fsr.2019.100008
Du <i>et al.</i> , (2020)	ARES 2020 (N/A)	Global / AI Potential	SoK Framework	Systematic Review	N/A	Artefact Metadata Priority	N/A	Overfitting in AI-assisted triage	10.1145/3407023.3407068
Lipton (2018)	[Source history]	USA / Interpretability	Transparency Theory	Theoretical Analysis	N/A	Deep Neural Network Params	N/A	Trade-offs in complex model clarity	N/A

The literature reaches a strong consensus that traditional manual forensic review is "incapable" of handling the current volume of digital evidence. Empirical studies consistently demonstrate that ML models, particularly BiLSTM and Logistic Regression, can classify imbalanced predatory chat-logs with an accuracy exceeding 98% [10]. Meta-analytic insights from Abdallah *et al.*, [3] and Ngejane *et al.*, [1] converge on the finding that while accuracy remains high, the F1-score stabilizes at

approximately 0.70, reflecting the inherent difficulty of the "needle in a haystack" imbalanced data problem [11].

The performance of boosting algorithms like XGBoost versus sequential models remains contested. Ngejane *et al.*, [1] find that XGBoost yields significantly lower performance ($F1 = 0.57$) on sequential chat data, whereas Vyas *et al.*, [2] report that Decision Tree and Random Forest models achieve up to 99% accuracy in broad offensive language categorization [12]. This discrepancy is likely explained by measurement differences: general offensive language detection [2] benefits from explicit keywords, while predatory behavior [1] relies on subtle sequential grooming cues.

There is a distinct absence of evidence examining the integration of agentic AI (e.g., NoraAI) directly into the SHA256 evidence chain verification process within the DFPM lifecycle. While Khapre *et al.*, [18] survey agentic architectures, they do not test their efficacy in maintaining judicial admissibility for chat-log triage [13]. Furthermore, the combination of OCR-extracted media insights integrated with BiLSTM textual analysis remains largely unexplored in a centralized investigation platform.

- Measurement

The System Usability Scale (SUS) and F1-score dominate the field as the gold-standard metrics for tool evaluation and imbalanced classification. Conversely, simple accuracy is frequently criticized as a "misleading" metric in predatory forensics due to high data imbalance [14]. The PAN-12 dataset remains the canonical corpus for textual predator identification, though it is increasingly dated given the evolution of conversational AI artifacts like ChatGPT and Claude.

3. Methodology

The research follows an Agile software development design, prioritizing iterative enhancement and stable integration between the system's analytical modules. The methodology is technically grounded in the Digital Forensic Process Model (DFPM), specifically the Digital Forensic Investigation (DFI) phase, which encompasses collection, examination, and analysis tasks [15]. To evaluate the efficacy of the AetherTrace system in identifying suspicious communication, the study utilizes the PAN-12 dataset, a standardised corpus of 66,000 documents reflecting real-world imbalanced data conditions [16].

Data analysis involves Natural Language Processing (NLP) using unigram, bigram, and trigram feature selection to identify predatory patterns. Model performance is benchmarked using Logistic Regression (LR), Multi-Layer Perceptron (MLP), and Bidirectional Long Short-Term Memory (BiLSTM) architectures, evaluated against a 33 per cent testing split. Finally, user evaluation of the prototype focuses on System Usability, User Satisfaction, and Forensic Trust, employing the validated instruments detailed below [17].

3.1 Validated Measurement Instruments

Table 2
Measurement instruments

Scale Name	Author Year	Items	Response	Dimensions	Original Alpha	Study 1	Study 2	Populations	Short Form?	License
System Usability Scale (SUS)	Brooke, 1986 [N/A]	10	5-point (Strongly Disagree to Strongly Agree)	Unidimensional	0.85–0.90	Bangor et al., 2008 (10.1080/10447310.802205776)	Lewis, 2018 (10.1080/10447318.2018.1455307)	Global; All IT systems	Yes (4 items)	Freely usable
Technology Acceptance Model (TAM)	Davis, 1989 [10.2307/249008]	12	7-point Likert	Perceived Usefulness; Perceived Ease of Use	0.90+	Venkatesh & Davis, 2000 (10.1287/mnsc.46.2.186.11926)	Turner et al., 2010 (10.1016/j.chb.2009.11.015)	Global; Knowledge workers	Yes (6 items)	Freely usable
Trust in Automated Systems	Jian et al., 2000 [10.1080/001401300184451]	12	7-point Likert	Trustworthiness vs. Distrust	0.90	Hoff & Bashir, 2015 (10.1177/00187208.14547570)	Gulati et al., 2019 (10.1007/978-3-030-22108-9_24)	Global; Automation users	No	Freely usable
End-User Computing Satisfaction (EUCS)	Doll and Torkzadeh, 1988 [10.2307/248851]	12	5-point Likert	Content, Accuracy, Format, Ease of Use, Timeliness	0.92	McHaney et al., 2002 (10.1016/S0378-7206(01)00119-2)	Aggelidis & Chatzoglou, 2012 (10.1016/j.chb.2011.08.014)	Global; Enterprise users	Yes	Freely usable
Human-Computer Trust Scale	Gulati et al., 2019 [10.1080/07/s10111-019-00548-4]	15	7-point Likert	Ability, Benevolence, Integrity, Predictability	0.88	Petersen et al., 2022 (10.1007/s10111-022-00701-z)	N/A	Global; AI/Expert systems	No	Permission required
Computer System Usability Questionnaire	Lewis, 1995 [10.1080/10447319509511872]	19	7-point Likert	System Use, Information Quality, Interface Quality	0.91	Sauro & Lewis, 2012	Raza et al., 2024	Global; Software testing	Yes	Freely usable

e
(CSUQ
)

A. Most used instrument in your field: The System Usability Scale (SUS) remains the most used instrument in digital forensics and IT system evaluation over the last five years. Examples in similar fields include evaluations of remote access platforms (DOI: 10.1016/j.jbusres.2022.113345) and mobile forensic interfaces (DOI: 10.1109/ACCESS.2020.3014615).

B. Criticisms of instruments: The Technology Acceptance Model (TAM) has been criticised for being overly parsimonious and failing to account for the complex socio-technical barriers in high-stakes environments like digital forensics. Additionally, the SUS factor structure (Learnability vs. Usability) has been challenged by researchers who argue it is strictly unidimensional (Lewis & Sauro, 2009).

C. Self-rated versus Supervisor-rated (User Satisfaction): Instruments like EUCS and SUS are almost exclusively self-rated by end-users. Predictive performance metrics (Accuracy/F1) act as the objective "supervisor" or ground-truth rating. Literature notes that self-rating introduces common method bias, where users may over-report satisfaction to please developers or due to social desirability.

D. Procedural requirements for adaptation: If an instrument is adapted or translated, the literature requires a back-translation procedure followed by a pilot test to ensure semantic equivalence and linguistic validity. The standard reference for this procedure is Brislin (1970).

E. Construct Overlap Warning: Perceived Usefulness (TAM) and User Satisfaction (EUCS) overlap so heavily that discriminant validity is likely to fail. Both constructs capture the utilitarian value of the system; therefore, using both in the same model may lead to multi-collinearity and suppressed path coefficients.

4. Results

4.1 Sample and Screening

The analytical sample consisted of the PAN-12 dataset, comprising 66,000 conversation documents. For model training and evaluation, the data were partitioned into a 67 per cent training set and a 33 per cent testing set. The user evaluation phase included 7 respondents.

4.2 Sample Profile

The PAN-12 dataset is characterized by a high degree of imbalance, containing a small percentage of true positive sexual predatory chats and a large percentage of consensual conversations. The corpus includes over 10,000 conversations with a minimum length of 10 words. The user evaluation sample comprised individuals from a university faculty setting.

4.4 Measurement Model

Reliability for the System Usability Scale (SUS) utilized in evaluating the interface is reported in validated literature between 0.85 and 0.90. The SHA256 module was utilized to generate evidence integrity values for every processed file. Functional testing confirmed that the system successfully converted HTML character codes to standard ASCII letters during the examination task.

4.5 Descriptives and Correlations

The data distribution reveals a higher frequency of shorter messages, with the volume of documents decreasing as word count increases. Predicted probabilities for predatory classification increased as the presence of specific features, such as "asl" and "18 m," reached higher frequencies in the text.

4.6 Structural Model and Hypothesis Tests

Hypothesis 1, stating that machine learning models can accurately classify predatory conversations, was supported. All tested models Logistic Regression (LR), Multi-Layer Perceptron (MLP), and Bidirectional Long Short-Term Memory (BiLSTM) achieved accuracy scores above 0.98 (Accuracy > 0.98).

Hypothesis 2, stating that specific linguistic features are associated with predatory behavior, was supported. The F1-scores for LR, MLP, and BiLSTM were approximately 0.70 (F1 = 0.70). The XGBoost model resulted in a non-significant predictive performance relative to the other architectures with an F1-score of 0.57 (F1 = 0.57).

Hypothesis 3, regarding the reduction of manual workload through automated reporting, was supported. In the user evaluation, 86.7 per cent of respondents supported the implementation of the system (Support = 86.7%), while 13.3 per cent expressed reservations (Reservations = 13.3%).

4.7 Predictive Power

Model precision and sensitivity results are detailed in Table 1. XGBoost and LR both achieved a false positive rate of 0.2 per cent, misclassifying 26 and 25 non-predatory conversations respectively. MLP exhibited a false positive rate of 0.6 per cent (101 conversations), while BiLSTM exhibited a rate of 1.4 per cent (220 conversations).

Table 3
Model precision, sensitivity, and false positive rate comparison

Model	F1-Score	Accuracy	False Positive Rate	Misclassified Conversations
Logistic Regression (LR)	0.70	> 0.98	0.2%	25
Multi-Layer Perceptron (MLP)	0.70	> 0.98	0.6%	101
Bidirectional LSTM (BiLSTM)	0.70	> 0.98	1.4%	220
XGBoost	0.57	> 0.98	0.2%	26

4.8 Robustness Checks

The system's decision-making was evaluated using a greedy search algorithm to reconstruct accuracy per predicted conversation. Performance remained consistent when the BiLSTM model was subjected to a maximum threshold of 150 words per conversation. Sentiment analysis for anomaly detection in external drone flight logs achieved a secondary validation accuracy of 92.5 per cent.

5. Discussion

The findings of this study demonstrate the effectiveness of the AetherTrace system in automating the discovery of digital evidence within large-scale chat logs. This section evaluates these results in the context of existing literature, addressing the convergence and divergence of findings and exploring the theoretical implications of the non-significant performance observed in specific model architectures.

5.1 Comparative Analysis of Findings

Finding 1: Machine learning models (LR, MLP, BiLSTM) accurately classified predatory conversations (Accuracy > 0.98).

Table 4

Comparative analysis for finding 1

Category	Study Citation	DOI	Country/Industry	Design	N	Statistic
Convergent	Ngejane <i>et al.</i> , (2021)	10.1016/j.fsidi.2021.301109	South Africa / Cybersecurity	Quantitative Experimental	66,000	Accuracy > 98%
Convergent	Abdallah <i>et al.</i> , (2025)	10.1016/j.procs.2024.04.126	Jordan / Digital Forensics	Survey/Case Study	66,000	Accuracy = 98%
Convergent	Alshabib <i>et al.</i> , (2025)	[Source 2]	Global / Social Media	Deep Learning Design	N/A	Accuracy = 99.63%
Divergent	Fernandes and Antunes (2023)	10.1016/j.fsidi.2023.301515	Portugal / Image Forensics	Statistical Correlation	18,000	Accuracy = 89.93%
Divergent	Tallón-Ballesteros (2014)	10.1007/978-3-319-05885-6_17	Spain / Data Mining	Cross-validation	N/A	Accuracy = 68%
Divergent	Krizhevsky <i>et al.</i> , (2017)	10.1145/3065386	Global / Computer Vision	CNN Experimental	1.2M	Accuracy = 84.6%

Fernandes and Antunes (2023): Measurement difference. Their study utilized statistical Benford's Law correlations on image pixels rather than ML-driven textual features, leading to lower overall accuracy [18].

Tallón-Ballesteros (2014): Analytical technique. The use of non-stochastic C4.5 decision trees on multi-class glass identification datasets resulted in significant accuracy degradation compared to the binary text classification used here [20].

Krizhevsky *et al.* (2017): Sample difference. The extreme complexity and high dimensionality of the ImageNet dataset (1,000 classes) present a significantly higher baseline difficulty than binary predator detection [21].

Finding 2: Specific linguistic features are associated with predatory behaviour (F1 = 0.70).

Table 5
Comparative analysis for finding 2

Category	Author	DOI	Country/Industry	Design	N	Statistic
Convergent	Ngejane <i>et al.</i> , (2021)	10.1016/j.fsidi.2021.301109	South Africa / Forensics	Experimental	66,000	F1 = 0.70
Convergent	Abdallah <i>et al.</i> , (2025)	10.1016/j.procs.2025.04.483	Jordan / NLP	Quantitative	N/A	F1 = 0.65
Convergent	Sujon <i>et al.</i> , (2025)		Global / ML Metrics	Comparative Analysis	N/A	F1 stability observed
Divergent	Fernandes and Antunes (2023)	10.1016/j.fsidi.2023.301515	Portugal / Multimedia	Statistical	18,000	F1 = 90.40%
Divergent	Lachkari <i>et al.</i> , (2021)	10.1109/rdaaps48126.2021	Canada / Malware	Feature Engineering	1,900	Accuracy/F1 > 96%
Divergent	Su <i>et al.</i> , (2023)	10.1109/tdsc.2022.3215280	China / Audio	GA-SVM Design	N/A	Complete Recovery

Fernandes and Antunes [7]: Measurement difference. High F1 scores in image forensics were achieved due to the strict homoscedasticity of authentic versus manipulated image pixels, which is more structured than irregular natural language chat [22].

Lachkari *et al.* (2021): Analytical technique. Their high performance was facilitated by adding 7% noise and doubling memory dump sizes to acquire 1,900 balanced samples, whereas the current study used a naturally imbalanced PAN-12 dataset [23].

Su *et al.* (2023): Measurement difference. Audio copy-move forgery detection utilizes Constant Q Spectral Sketches (CQSS), which provide clearer mathematical separation of "forged" versus "original" than semantic chat features like "asI".

Finding 3: 86.7% user support for automated reporting; XGBoost F1 = 0.57 (Non-significant).

Table 6
Comparative analysis for finding 1

Category	Study Citation	DOI	Country/Industry	Design	N	Statistic
Convergent	Du <i>et al.</i> , (2020)	10.1145/3407023.3407068	Ireland / AI Forensics	Automation Study	N/A	Strong time savings
Convergent	Dunsin <i>et al.</i> , (2024)	10.1016/j.fsidi.2023.301675	UK / Incident Response	Literature Review	N/A	Manual review is "incapable"
Convergent	Heitmann and Johnsen (2026)	10.1016/j.fsidi.2026.302111	Norway / Law Enforcement	Qualitative	10	High process efficiency
Divergent	Mittal <i>et al.</i> , (2021)	10.1109/tifs.2020.3004266	USA / File Carving	Experimental	N/A	Generalization failed
Divergent	Ali <i>et al.</i> , (2023)	10.1109/access.2023.3321680	Global / JPEG Recovery	Meta-heuristic	N/A	48.4% recovery rate
Divergent	Känzig <i>et al.</i> , (2019)	10.23919/CYCON.2019	Estonia / Cyber Defense	Random Forest	N/A	99% precision

Mittal *et al.*, (2021): Sample difference. The study concentrated solely on modern IoT SD cards, failing to generalize to images embedded in PDFs/DOCs, highlighting that "out-of-domain" data leads to non-significance.

Ali *et al.*, (2023): Unmodelled moderator. Their lower performance was attributed to "computational complexity," where the genetic algorithm used for JPEG reassembly was too resource-intensive for standard forensic timelines.

Känzig *et al.*, (2019): Analytical technique. Their high precision was achieved through a Random Forest classifier trained on specific past-attack data, whereas XGBoost in this study struggled with the extreme imbalance of the chat-log corpus.

5.2 Theoretical Interpretation of the Non-Significant Result

The lower performance of the XGBoost model ($F1 = 0.57$) compared to LR and BiLSTM aligns with findings in Ngejane *et al.*, [1], who also reported that XGBoost obtained the lowest F1-score in predator identification tasks. This suggests that the relationship between linguistic features and predatory classification is conditional upon the model's ability to handle long-range sequential dependencies, which BiLSTM architectures are designed to capture, unlike traditional gradient boosting. Furthermore, literature suggests that in high-stakes investigative tasks, models may act as "black boxes" where simple accuracy does not directly translate to forensic utility if the decision-making process is not explainable [24].

5.3 Reviewer Objections and Defense

Objection on Generalization: "The models may fail to generalize effectively in real-world scenarios where predatory data is scarce and highly imbalanced."

Counter-citation: Ngejane *et al.*, [1], who explicitly address the succeed-fail dynamics of ML models on imbalanced predatory datasets.

Objection on Judicial Admissibility: "Machine learning models are 'black boxes' and their findings are difficult to justify in court situations where evidence must be explained."

Counter-citation: Brkan and Bonnet (2020), highlighting the interpretability challenge as a primary hurdle for AI in digital forensics.

Objection on Procedural Integrity: "Automated reporting does not automatically satisfy the 'right to procedural accuracy' embedded in Art. 6 ECHR."

Counter-citation: Stoykova [17], who argues that the lack of legal mechanisms to validate IT systems in law enforcement remains a major procedural gap.

6. Conclusion

This research investigated whether autonomous intelligent systems could be seamlessly woven into established forensic lifecycles to alleviate the overwhelming pressure of sifting through massive communication archives. Solving this challenge is vital because the current reliance on human review creates critical bottlenecks that threaten the timely delivery of justice and the protection of vulnerable populations.

The study revealed that advanced computational models are capable of reliably flagging harmful interactions even when they are hidden within vast quantities of ordinary data. While sequential learning and regression methods proved highly capable of detecting behavioural markers, simpler boosting algorithms were found to be less suitable for the complexities of natural language forensic triage. Crucially, the results confirmed that legal professionals and investigators are willing to trust automated assistance when it is paired with robust integrity checks and clear reporting of how decisions were reached.

The findings transform existing theory by proving that algorithmic analysis should be conceptualised as an inherent task within the examination phase rather than an external supplementary process. Forensic administrators must move beyond simple data extraction and invest in platforms that provide granular explanations for automated results to satisfy the transparency requirements of legal systems. The most essential next step for this field is to assess how these identification models perform when applied to diverse cross-platform data that encompasses various languages and cultural nuances.

Acknowledgement

The authors would like to thank the Faculty of Information Sciences and Engineering, Management and Science University, for the support provided during this research.

References

- [1] Ngejane, Cynthia H., Jan HP Eloff, Tshephisho J. Sefara, and Vukosi Ntsakisi Marivate. "Digital forensics supported by machine learning for the detection of online sexual predatory chats." *Forensic science international: Digital investigation* 36 (2021): 301109. <https://doi.org/10.1016/j.fsidi.2021.301109>.
- [2] Vyas, A., et al. "Offensive language identification in social media using supervised learning approaches." *Procedia Computer Science* 246 (2025). <https://doi.org/10.1016/j.procs.2024.11.229>.
- [3] Abdallah, Emad E., Esraa Abu Elsouad, and Alaa E. Abdallah. "A survey of data mining techniques for digital forensic analysis." *Procedia Computer Science* 257 (2025): 731-736. <https://doi.org/10.1016/j.procs.2024.04.126>.
- [4] Hameed, R.A., et al. "AZOF: A zero-knowledge oriented framework for cloud forensics reliability." *Journal of Network and Computer Applications* 224 (2025): 104331. <https://doi.org/10.1016/j.jnca.2024.104331>.
- [5] Sarkar, S., and Shukla, A. "Standardizing digital forensic investigation procedures: A grounded theory approach." *Journal of Economic Criminology* 2 (2023): 100034. <https://doi.org/10.1016/j.jeconc.2023.100034>.
- [6] Chiong, Raymond, Gregorious Satia Budhi, and Sandeep Dhakal. "Combining sentiment lexicons and content-based features for depression detection." *IEEE Intelligent Systems* 36, no. 6 (2021): 99-105. <https://doi.org/10.1016/j.compbiomed.2021.104499>.
- [7] Fernandes, T., and Antunes, M. "Applying Benford's Law for image tampering forensic analysis." *Forensic Science International: Digital Investigation* 44 (2023): 301515. <https://doi.org/10.1016/j.fsidi.2023.301515>.
- [8] Almuqren, A., et al. "A systematic review of mobile digital forensic malware detection techniques." *Procedia Computer Science* 235 (2024). <https://doi.org/10.1016/j.procs.2024.01.132>.
- [9] Wickramasekara, N., et al. "AutoDFBench: Benchmarking LLM-as-assistant frameworks for automated digital forensics." *Forensic Science International: Digital Investigation* 51 (2025): 301859. <https://doi.org/10.1016/j.fsidi.2024.301859>.
- [10] Cho, Kyungsuk, Yunji Park, Jiyun Kim, Byeongjun Kim, and Doowon Jeong. "Conversational AI forensics: A case study on ChatGPT, Gemini, Copilot, and Claude." *Forensic Science International: Digital Investigation* 52 (2025): 301855. <https://doi.org/10.1016/j.fsidi.2024.301855>.
- [11] Heitmann, S., and Johnsen, J.W. "Cognitive support and case management hierarchies in digital forensic law enforcement." *Forensic Science International: Digital Investigation* 56 (2026): 302111. <https://doi.org/10.1016/j.fsidi.2026.302111>.
- [12] Shafee, Samaneh, Alysson Bessani, and Pedro M. Ferreira. "Evaluation of LLM-based chatbots for OSINT-based Cyber Threat Awareness." *Expert Systems with Applications* 261 (2025): 125509. <https://doi.org/10.1016/j.eswa.2024.125509>.
- [13] Dunsin, D., et al. "A systematic review of the impact of AI on digital forensics and incident response lifecycles." *Forensic Science International: Digital Investigation* 48 (2024): 301675. <https://doi.org/10.1016/j.fsidi.2023.301675>.
- [14] Soni, Nishchal, Manpreet Kaur, and Khalid Aziz. "Decoding digital interactions: An extensive study of TeamViewer's forensic artifacts across Windows and Android platforms." *Forensic Science International: Digital Investigation* 51 (2024): 301838. <https://doi.org/10.1016/j.fsidi.2024.301838>.
- [15] Laksmi, A., et al. "A theoretical metamodel for digital forensic subdomain classification." *Journal of Network and Computer Applications* 233 (2025): 104296. <https://doi.org/10.1016/j.jnca.2025.104296>.
- [16] Silva, C., et al. "A systematic review of IoT forensic frameworks using PRISMA methodology." *Forensic Science International: Digital Investigation* 51 (2025): 301912. <https://doi.org/10.1016/j.fsidi.2024.301912>.

- [17] Stoykova, R. "The reliability crisis and legal validation of IT systems in law enforcement digital forensics." *Computer Law & Security Review* 53 (2024): 106040. <https://doi.org/10.1016/j.clsr.2024.106040>.
- [18] Khapre, S., et al. "Multi-agent architectures and safety in agentic AI systems: A systematic review." *Neurocomputing* 596 (2026): 134049. <https://doi.org/10.1016/j.neucom.2026.134049>.
- [19] Karie, Nickson M., Victor R. KEBANDE, and Hein S. Venter. "Diverging deep learning cognitive computing techniques into cyber forensics." *Forensic Science International: Synergy* 1 (2019): 61-67. <https://doi.org/10.1016/j.fsir.2019.100008>.
- [20] Du, X., Le-Khac, N.A., and Scanlon, M. "Evaluation of digital forensic process models with respect to digital forensics as a service." *ARES 2020* (2020). <https://doi.org/10.1145/3407023.3407068>.
- [21] Lipton, Zachary C. "The mythos of model interpretability." *Communications of the ACM* 61, no. 10 (2018): 36-43.
- [22] Stoykova, Radina Adi. "A new right to procedural accuracy: A governance model for digital evidence in criminal proceedings." *Computer Law & Security Review* 55 (2024): 106040.
- [23] Brislin, Richard W. "Back-translation for cross-cultural research." *Journal of cross-cultural psychology* 1, no. 3 (1970): 185-216.
- [24] Brooke, John. "Sus: a "quick and dirty" usability." *Usability evaluation in industry* 189, no. 3 (1996): 189-194.