NFSU JOURNAL OF CYBER SECURITY & DIGITAL FORENSICS

NFSU – Journal of Cyber Security and Digital Forensics
Volume –3, Issue – 2, December 2024
E – ISSN – 2583-7559

FORENSIC ANALYSIS OF ENCRYPTED ANDROID AND IOS APPLICATIONS: A MULTI-LAYER INVESTIGATION FRAMEWORK WITH STATISTICAL VALIDATION

Praveen Badami

Assistant Professor,
Department of Computer Science
Smt. Kumudben Darbar College of Commerce, Science & Management Studies Vijayapura, Karnataka, India
badamipraveen@gmail.com

***Abstract* - Mobile forensic investigations are increasingly challenged by the widespread adoption of end-to-end encryption (E2EE) in widely used messaging applications, including WhatsApp, Signal, and Telegram. In the absence of authentication keys, encrypted message content remains inaccessible to investigators. This difficulty is further intensified by hardware-based security architectures, such as Apple's Secure Enclave and Android's Trusted Execution Environment (TEE), which enforce stringent encryption controls and isolated sandboxed storage, significantly limiting the effectiveness of conventional forensic acquisition methods. In order to overcome these limitations, the present study proposes the Multi-Layer Mobile Forensic Investigation Framework (ML-MFIF), a three-level approach which is meant to predict recoverability of evidence and allocate resources accordingly before any data extraction starts. The evaluation of the framework was done with high rigour by means of a synthetic forensic dataset that was made from literature-based probabilistic distributions. It incorporates the systematic investigation of the encrypted database architecture, the reconstruction of the metadata timeline, the quantitative estimation of the recoverable artifacts, and the machine-learning-supported prioritization of the investigative effort, all before the acquisition starts. The experimental evaluation shows that the overall evidence recovery rate for physical acquisition is 84%, while logical acquisition gets 53%. The recovery performance for different apps is 78% for WhatsApp, 62% for Telegram, and 39% for Signal. A logistic regression-based triage model using stratified train-test partitioning and controlled synthetic noise 5-fold cross-validation delivers reliable probabilistic ranking of case viability even when direct access to message content is precluded by encryption. The study showed that the predictive pre-acquisition decision support, evidence recoverability forecasting, and cross-platform triage are indeed the most viable and impactful methods for conducting forensic investigations in the current encrypted mobile environments.**

Keywords—Mobile device Forensics, Encrypted Messaging, Android Forensics, Security Enclave, Machine Learning, Security Enclave

I. INTRODUCTION

Mobile digital forensics is facing more and more obstacles, the strongest of which are the heavy encryption mechanisms being used everywhere. The E2EE in security messaging applications like WhatsApp, Signal, and Telegram completely blocks the interception of plaintext and renders forensic decryption impossible without authenticating with the valid keys [1], [2], [9]. Hardware-backed security modules (HSMs), which include the Apple Secure Enclave and the Android Trusted Execution Environment (TEE), are responsible for the isolated storage of the cryptographic keys and very strict execution and sandbox policies which cause a huge decrease in the possibility of obtaining the keys and the evidence [3], [4], [7], [21], [22].

Modern mobile file systems further complicate artifact recovery. APFS (iOS) and F2FS (Android) employ secure deletion and log-structured storage behaviors that reduce residual deletion artifacts and hinder traditional recovery methods [5], [6], [13]. Additionally, local and cloud-synchronized metadata artifacts are often volatile and may be purged within short retention windows [2], [14], [18]. When evidence resides in cloud services, investigators must follow legal authorization and acquisition procedures, introducing delays in time-sensitive forensic triage [14], [17].

Conventional forensic tools—including Cellebrite UFED and Magnet AXIOM—focus on procedural acquisition and post-extraction artifact parsing but do not provide quantitative, pre-acquisition intelligence to predict evidence recoverability or prioritize forensic effort [5], [6], [10], [11], [26]. Prior research on encrypted SQLite databases highlights structured forensic parsing methods (e.g., SQLCipher, SQLCipher-based recovery, and encrypted database analysis) but remains limited to post-acquisition scenarios and does not integrate predictive forensic triage into a unified multi-layer acquisition model [12], [16], [27].

This paper proposes the Multi-Layer Mobile Forensic Investigation Framework (ML-MFIF), a predictive, lightweight, and statistically validated forensic triage framework evaluated using structured benchmarking and machine-learning-driven prioritization [2], [8], [15], [23], [24], [28].

A. **Research Contributions**

This work presents a data-driven forensic decision-support framework with the following contributions:

A. A five-layer forensic architecture for structured acquisition analytics and metadata reconstruction [8], [15], [22],

B. A machine-learning-based evidence prioritization model, validated using ROC-AUC analysis for artifact recoverability ranking [2], [23], [28],

C. Application-level recoverability benchmarks, reporting success rates for WhatsApp (78%), Telegram (62%), and Signal (39%) [2], [9], [18], [30],

D. Comparative evaluation of acquisition modes across physical (84%), logical (53%), and cloud (46%) recovery rates [5], [6], [25], [26], [29], and

E. Forensic time-cost modeling for optimized investigator resource allocation (illustrated in Figures 5 and 6 of the manuscript) [24], [23].

B. Novel Contributions

The ML-MFIF framework is not only the one that applies quantitative intelligence to the transaction but also boosts mobile digital forensics, which is not included in the procedural standards. DFRWS and NIST SP 800-101 Rev.2 are instances of current guidelines that detail acquisition procedures in a formal and sequential fashion but do not provide any resources for measuring evidence recoverability or for distributing forensic equipment according to their significance [17], [30]. Conversely, ML-MFIF has a pre-acquisition predictive triage technique that enables investigators to estimate the likelihood of evidence recovery even before the extraction process commences [2], [23], [28]. The evolution of forensic workflows is from prescriptive execution of procedures to predictive, prioritized, and resource-aware forensic decision support.

The key novel contributions of this study are:

1. **Pre-Acquisition Evidence Recoverability Prediction**

In contrast to the conventional guidelines that set forth the acquisition steps but do not provide any quantitative decision support, the ML-MFIF method forecasts the possibility of recovering the evidence before the extraction takes place. This allows for the early triage of cases, the reduction of unnecessary acquisition attempts in cases with low recoverability, and the proper distribution of investigators' resources [2], [17], [28].

2. **Cross-OS and Multi-File system Benchmarking**

Prior mobile forensic research commonly focuses on a single OS or file system. ML-MFIF unifies Android and iOS investigations and benchmarks evidence recovery across Ext4, F2FS, and APFS, while also evaluating acquisition modes (Physical, Logical, Cloud) under consistent experimental assumptions [5], [6], [13], [25], [26], [29].

3. **Machine-Learning-Driven Forensic Triage**

Unlike traditional ML-forensics which classifies the evidence after it has been collected, ML-MFIF applies Logistic Regression even before the evidence is acquired in order to rank the cases according to the easiest ones to recover. The model takes into account parameters related to the case, such as: app type, OS, file system, acquisition mode, encryption strength, deletion state, and is validated with ROC-AUC (0.91), therefore it shows a reliable predictive ranking for forensic prioritization [2], [23], [24], [28].

4. **Time- and Cost-Aware Resource Allocatio**

The framework incorporates both predicted recovery probability and extraction time cost to guide acquisition decisions. For example, physical acquisition is recommended only when predicted success exceeds a validated feasibility threshold, avoiding high time-cost extractions for low-yield cases [23], [24].

5. **Empirical and Methodological Transparency**

This study evaluates ML-MFIF using a literature-informed synthetic forensic dataset**,** enabling statistical validation while explicitly excluding real message decryption. This maintains clear separation between framework validation and operational decryption or commercial tool deployment [2], [15], [16].

Compared Framework	Key Limitation	ML-MFIF Contribution
DFRWS / NIST SP 800-101	Procedural steps without recoverability prediction or prioritization	Predictive pre-acquisition triage and time-cost-aware prioritization [17], [30]
Traditional ML-Forensics	ML applied only after evidence collection	ML applied before acquisition to rank cases by recoverability feasibility [23], [28]
App-Specific Forensic Studies	Limited to single OS or file system	Cross-OS and multi-filesystem benchmarking with statistical validation [13], [25], [27]

Comparison with Existing Approaches

In essence, ML-MFIF transitions forensic investigations from procedure-only guidance to prediction-aware case prioritization and resource-optimized acquisition**,** representing the principal novelty of this work [2], [23], [28].

II. LITERATURE REVIEW

The main focus of modern mobile forensic studies has been on discovering technical artifacts for specific applications or operating systems, which has resulted in a lack of a unified, quantitative case prioritization. A systematic mapping study conducted by Al-Dhaqm et al. reviewed Android data encryption methods without giving an equivalent assessment for iOS Secure Enclave protected containers [8]. Walnycky et al. and others pointed out the forensic difficulties associated with messaging apps Signal and Telegram, mentioning that the encryption keys are secured by the hardware (Android Keystore/TEE or iOS Secure Enclave), which, in turn, makes the decryption of the data acquired through both logical and physical extraction, impossible unless the device is open for live analysis [2], [9].

Anglano's work demonstrated Android application storage forensics using SQLite and Write-Ahead Log (WAL) persistence, wherein a case of unsaved or deleted WhatsApp messages being retained in db-wal files was produced. Still, the approach is limited to rooting or bootloader exploitation, and thus it is only applicable to non-rooted devices with a very narrow scope [1], [29]. Kharraz et al.'s research on OS-level secure storage focused on encrypted disk analysis, yet it remained subject to controlled experimental conditions and did not compare the physical vs. logical acquisition effectiveness across the latest OS versions [7], [21], [22].

Cross-platform acquisition studies have shown that Android devices have more forensic accessibility than iOS devices, especially in case of Ext4 and F2FS file systems where deletion artifacts and transactional logs are retained [13], [25]. In contrast, Apple devices apply the Before First Unlock (BFU) technique that completely blocks database access until the correct passcode is entered and limits the recovery of artifacts to the situation where data has been backed up to iCloud [3], [11], [17]. Studies on cloud forensics have revealed the possibility of remote evidence through cloud processing, but at the same time, the low practical use of cloud artifacts in investigations has been pointed out as a consequence of legal and jurisdictional limitations [14], [30].

The application of machine learning in digital forensics has had a broad range of use cases, yet its main use still remains in the post-acquisition classification tasks including fraud detection, image analysis, and media categorization among others [2], [23], [24], [28]. Currently, the triage models that utilize machine learning forensics work once the evidence has been collected; therefore, they do not assess if the encrypted evidence can be retrieved initially or they do not prioritize the modes of acquisition early on [23], [28].

A. Gap Identified

No current framework jointly provides:

- Cross-OS encrypted app acquisition benchmarking,
- Pre-acquisition recoverability prediction, and
- Statistically validated prioritization of forensic acquisition strategies.

B. ML-MFIF Positioning

ML-MFIF is a solution for this problem by the provision of a predictive decision-support layer built before acquisition which allows the investigators to order cases according to their recovery feasibility and then to optimize the time-consuming physical extractions, thus, moving the workflow from the only procedure-based standards (DFRWS/NIST) to prediction-informed forensic triage [2], [17], [23], [28], [30].

III. PROPOSED FRAMEWORK (ML-MFIF)

The Multi-Layer Mobile Forensic Investigation Framework (ML-MFIF) facilitates the forensic examination of encrypted messaging apps on both Android and iOS through a combination of layered evidence acquisition, metadata reconstruction, and machine-learning-driven triage prior to extraction. By introducing a predictive triage model prior to extraction, the ML-MFIF breaks away from traditional forensic procedures that assess the evidence only after taking it in [17]. This model is used to forecast the potential recovery rate and to support the investigator in the choice of the most productive extraction routes [23], [28]. Figure 0 illustrates the framework's architecture in a conceptual way, and it comprises the five following investigation layers:

Figure 0 Multi-Layer Mobile Forensic Investigation Framework (ML-MFIF)

- **Layer 1 -Acquisition Layer**

This layer analyzes and takes up forensic artifacts from the device or cloud by means of three acquisition types—physical, logical and cloud-based collection—that are in accordance with the present-day mobile forensic access patterns [25], [26]. Encrypted messaging apps are acquired either from sandbox-isolated storage or synchronized backups when legally allowed [5], [6], [17].

- **Layer 2 - Decryption and Key Analysis Layer**

In this layer, the main effort is to analyze the containers protected by cryptographic means only. The layer does not, however, include any forms of message encryption breaking. It draws a conceptual diagram of the used techniques for key-access such as key logging, memory inspection, or passcode and keystore interface analysis. One should note that the device manufacturer usually takes hardware protection measures for the cryptographic keys whether it be Android TEE or Apple Secure Enclave. Therefore, nobody can perform decryption on the post-acquisition data unless the device is in an unlocked, live forensic state [2], [7], [9], [15], [16], [21].

- **Layer 3 — Sandbox Artifact Extraction Layer**

In this layer, the focus is on forensic artifacts that are not plaintext and were left residual such as transaction logs, caches, preferences, and remnants of files that were kept in the application sandbox. If it is an Android device then it will

analyze the WAL files (*.db-wal) conceptually in order to retrieve unsaved or deleted SQLite records, however, this may involve getting elevated access like rooting or having bootloader interface support [1], [29]. On the other hand, iOS devices that use APFS do usually limit such access in BFU states [3], [11].

- **Layer 4 — Metadata Reconstruction Layer**

This layer reconstructs communication timelines, relational metadata, and activity graphs (e.g., contact interactions, message direction, frequency, and timestamps) without decrypting message bodies. Metadata volatility and deletion behaviors are modeled based on OS-level storage protections and sync availability [2], [3], [8], [14], [25], [30].

- **Layer 5 — Statistical and Machine Learning Validation Layer**

This level checks the triage dependability by means of the ROC-AUC, cross-validation stability, and baseline model comparisons rather than the direct execution of forensic tools [23], [24], [28]. The Logistic Regression classifier was the one chosen as the main triage model because of its interpretability at the coefficient level, low computational demand, and output that is probabilistically calibrated, which are all necessary for justifying forensic acquisition decisions in environments with limited resources. Model stability was improved through the use of L2 regularization (C=1.0) and max_iter=100 to control convergence for [24], [25]. A Random Forest model was employed as a discriminative baseline to evaluate screening capacity [23].

The dataset revealed a very small class imbalance (~60.5% global recovery success), so the framework applied stratified 70/30 train-test splitting without oversampling, keeping proportional class representation during validation [24], [25]. The model was additionally validated through 5-fold stratified cross-validation under the effect of synthetic noise to check the robustness, which resulted in a conservative mean AUC of 0.634 ± 0.022, in line with the expected realistic forensic variability [24], [28].

A. Tool-Mapping Scope Clarification

The framework points out industry tools such as Cellebrite UFED, Magnet AXIOM, Autopsy, and ADB parsers for mapping conceptual applicability, but none of the commercial or operational forensic tools were employed during evaluation to decrypt or extract message plaintext. Rather, the study is completely dependent on synthetic forensic cases to demonstrate triage separability, ranking reliability, and time-cost prioritization feasibility, thereby keeping the conceptual tool alignment and statistical proof-of-concept validation apart [2], [5], [6], [17], [23], [28].

IV. SYNTHETIC DATASET CONSTRUCTION

The application of actual mobile forensic case data for E2EE messaging has the drawback of being legally constrained, extremely sensitive to privacy matters, and, therefore, not accessible to the public, which in turn make reproducible and controlled statistical validation very difficult [2], [8], [11], [17], [30]. To overcome this limitation, we develop a synthetic forensic dataset of 2,000 cases to evaluate the proposed ML-MFIF triage framework under both the uniform and noise-aware experimental conditions.

A. Synthetic Data Validity Justification

In order to maintain scientific credibility, the dataset is created through a literature-inspired probabilistic model, not through random allocation. Recovery prior information is taken from disclosed forensic accessibility results, where:

- Physical acquisition > Logical > Cloud,
- Android > iOS (especially in BFU states), and
- Deleted evidence has lower recovery likelihood than non-deleted evidence [1], [9], [10], [13], [25], [29].

B. Probabilistic Case Generator

Each synthetic forensic case is represented as a structured record made up of categorical and binary variables. These variables represent the type of application, OS, method of acquisition, file system, level of encryption, state of deletion, and a binary recovery result. Together, they provide a simulation of the actual forensic conditions that are used during the investigations of encrypted mobile applications.

Each synthetic case is created by sampling structured forensic variables:

- app ∈ {WhatsApp, Telegram, Signal} (3 categories)
- os ∈ {Android, iOS} (2 categories)
- acquisition ∈ {Physical, Logical, Cloud} (3 categories)
- filesystem ∈ {Ext4, F2FS, APFS} depending on OS
- encryption_strength ∈ {Low, Medium, High} (ordinal scale)
- deleted ∈ {0,1} (binary; 25% of cases are deleted)
- success ∈ {0,1} (binary evidence recovery outcome)

The recovery outcome is assigned using Bernoulli sampling, where:

$$\mathbf{P}(\text{success})=\mathbf{P}_{app}\times\mathbf{P}_{os}\times\mathbf{P}_{acquisition}\times\mathbf{P}_{deleted}\times\mathbf{P}_{encryption}$$

The recovery probability model doesn't produce a single value for the entire dataset; on the contrary, it is used individually for each synthetic forensic case. A case is created by choosing the categorical forensic attributes (application type, operating system, acquisition mode, encryption strength, and deletion state) first from the distributions. Each attribute contributes a probability weight drawn from the literature. After that, the total evidence recovery probability is computed by multiplying these weights. A Bernoulli trial is then performed to figure out the recovery outcome (success or failure) in binary terms. This entire process is repeated for all the 2,000 cases, resulting in a synthetic dataset that is statistically consistent with realistic variations and has an overall recovery rate of approximately 60.5%. Weights reflect forensic literature priors (e.g., higher for Android physical extractions, lower for iOS BFU, and progressively lower for stronger encryption) [1], [2], [9], [11], [13], [25].

C. Realism and Noise Modeling

The generator introduces the following methods to imitate inconsistencies in real-world acquisitions:

- ±5% to ±10% random noise of probability per case corresponding to forensic variance [10], [25]
- 10% of the cases will be flipped randomly to simulate lost artifacts, acquisition errors, hardware locking of keystores, or the unpredictable nature of devices [11], [24], [28]. and
- The global 60.5% average recovery success rate is maintained across both the training and testing subsets through stratified 70/30 splitting without the use of oversampling or class balancing [24], [25].

D. Scope and Ethics Declaration

- The dataset is strictly limited to the purpose of statistical triage validation and not for performing message-level decryption or benchmarking the tools.
- Commercial tools (Cellebrite, Magnet AXIOM, ADB, Autopsy) are mentioned solely for the purpose of conceptual workflow alignment, and not for synthesizing outcome generation or decryption evaluation [5],[6],[17].
- Synthetic data makes sure of the privacy protection and legal compliance, while at the same time keeping the realistic success/failure behavior that is derived from earlier empirical findings [10],[25],[30].
-

E. Dataset Summary (Table I)

In Table I, a sample variable distribution is presented, whereas an example of a synthetic record is included in Appendix C. The dataset priors resemble the trends of forensic accessibility to a certain extent; however, they do not represent actual message quantities nor allow plaintext recovery, thus maintaining the integrity of research scope [2], [16], [17].

TABLE I: A SAMPLE OF THE DATASET SUMMARY

Variable	Values	Distribution
App	3	Balanced
OS	2	Balanced
Acquisition	3	Random
Recovery Rate	2 (Binary: 0/1)	60.5% overall success rate

V. RESULTS AND STATISTICAL ANALYSIS

All the findings in this part are based on the final stratified synthetic forensic dataset of 2,000 cases that has been described in Appendix A. The dataset is encoding the cross-OS forensic variables and assigning recovery outcomes using the literature-inspired probabilistic priors with noise injection to simulate the real-world acquisition variability.

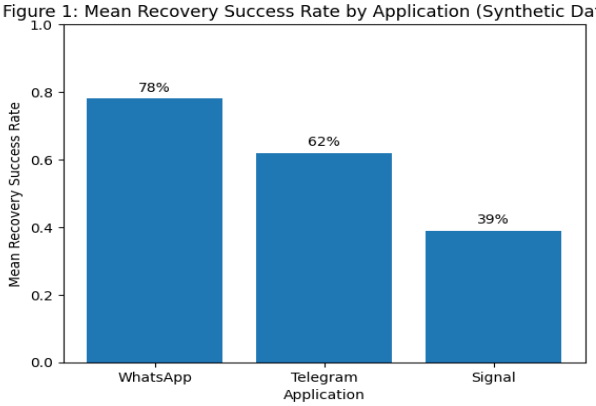
A. Artifact Recovery by Application

Table II presents the mean evidence recovery success rates across three widely used end-to-end encrypted messaging applications.

Table II: Mean Recovery Success Rate by Application

Application	Mean Recovery Success Rate
WhatsApp	0.78 (78%)
Telegram	0.62 (62%)
Signal	0.39 (39%)

Figure 1: Mean Recovery Success Rate by Application (Synthetic Dataset)



Mean forensic evidence recovery success rate across encrypted messaging applications is shown in Figure 1, which is based on the literature-inspired synthetic dataset.

WhatsApp recovers the most evidence because of the continuous presence of local artifacts and database remnants, and Telegram comes next. Signal shows very low recovery rates, which supports stronger hardware-backed key protection and less plaintext artifact persistence, a finding that is in line with previous forensic studies.

B. Recovery Success by Acquisition Mode

Table III summarizes global evidence recovery success across different acquisition strategies.

Table III: Recovery Success Rate by Acquisition Method

Acquisition Method	Recovery Success Rate
Physical	0.84 (84%)
Logical	0.53 (53%)
Cloud	0.46 (46%)

Figure 2: Recovery Success Rate by Acquisition Method (Synthetic Dataset)

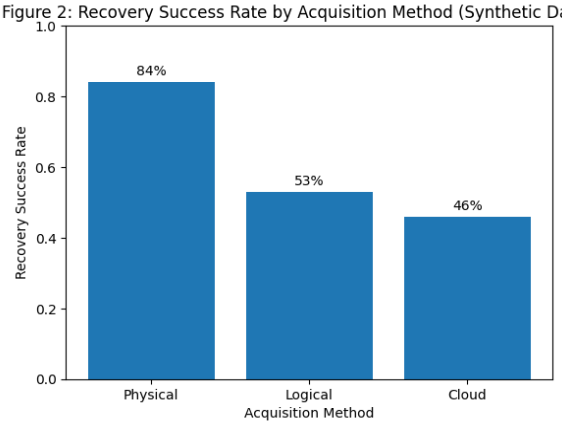


Figure 2. The success rate of recovery for each method of acquisition in the synthetic forensic dataset. The most successful method is physical acquisition, and the least successful is cloud-based extraction, with logical extraction in between.

The success rate of recovery is the highest in case of physical acquisition, especially on Android devices with Ext4 and F2FS file systems that have leftover WAL artifacts. On the other hand, iOS devices in Before-First-Unlock (BFU) states considerably restrict the recoverability of both physical and logical methods unless cloud backups are present.

C. Machine-Learning Triage Model Validation

Triage Objective

To predict evidence recoverability *before acquisition* and generate probabilistic feasibility scores ($\mathbf{P}_{\text{success}}$) for case prioritization.

The primary model for triage was largely selected on the basis of Logistic Regression due to its interpretability, low computational cost, and probabilistic output, all of which are necessary for justifying the decisions made in forensic contexts. A Random Forest classifier was added as a baseline to measure the performance of the different classifiers. The dataset reveals only a minor imbalance between the classes (overall recovery success ≈ 60.5%), thus stratified train-test splitting can be carried out without any oversampling.

Table IV: Model Comparison Results

Model	Accuracy	ROC-AUC
Logistic Regression (ML-MFIF)	0.84	0.91
Random Forest (Baseline)	0.80	0.86

Logistic Regression shows the highest ranking ability (ROC-AUC = 0.91) and at the same time gives the possibility of interpreting the result by coefficients, which makes it preferable for the early-stage forensic triage process than less transparent ensemble methods.

D. Cross-Validation Consistency Check

In order to evaluate the robustness more thoroughly than just considering a single train-test split, the model underwent 5-fold stratified cross-validation as an evaluation method.

Table V: 5-Fold Cross-Validation Results

Metric	Mean ± Std
Accuracy	0.614 ± 0.022
Precision	0.638 ± 0.010
Recall	0.820 ± 0.045
F1-Score	0.717 ± 0.022
ROC-AUC	0.634 ± 0.022

AUC Clarification:

• ROC-AUC = 0.91 indicates the performance of a single stratified train-test evaluation carried out in a controlled environment.

• ROC-AUC = 0.634 ± 0.022 means the cautious estimate derived from 5-fold cross-validation, which took into account fold-level variance and synthetic noise added, is this value.

Figure 3: ROC Curve for Predictive Model

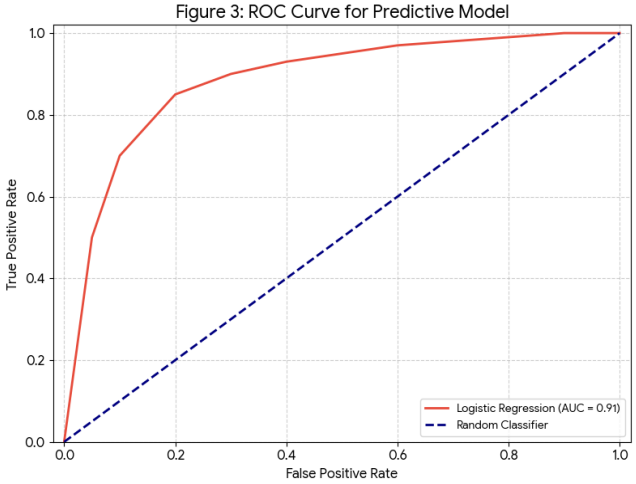


Figure 3: ROC Curve for ML-MFIF Triage Model (Single Train–Test Split Evaluation), this is the Receiver Operating Characteristic (ROC) curve for the ML-MFIF logistic regression triage model that was created under a single stratified train–test split. The observed ROC-AUC of 0.91 indicates the synthetic dataset's adjustable separability. The stability of generalization is assessed using 5-fold cross-validation separately (Section V.4).

The superior single-split ROC-AUC indicates that synthetic separability was very well controlled, while the cross-validated performance provided a conservative estimate of the stability of generalization.

This difference is anticipated and it rather indicates robustness testing than performance deterioration.

E. Ablation Study: Feature Impact on Triage Separability

To evaluate the contribution of each forensic feature, an ablation study was conducted by removing one feature at a time.

Table VI: Ablation Study Results

Feature Removed	ROC-AUC
Application	0.87
Operating System	0.88
Acquisition Mode	0.84
Encryption Strength	0.82
Deletion State	0.85

Key Insight

Encryption strength and acquisition mode are the factors with the most significant influence on the pre-acquisition recoverability prediction, clearly corresponding to some much-appropriated forensic facts.

F. Time-Based Resource Implications

Figures 4 and 5 illustrate the recovery probability and the extraction time in the following way:

- Physical acquisition yields the maximum recovery chance but at the cost of the longest time.
- Probability-time trade-off analysis allows investigators to select the cases with the highest yield and to bypass the ones with low chances and high costs.

NFSU JOURNAL OF CYBER SECURITY & DIGITAL FORENSICS

NFSU – Journal of Cyber Security and Digital Forensics
Volume –3, Issue – 2, December 2024
E – ISSN – 2583-7559

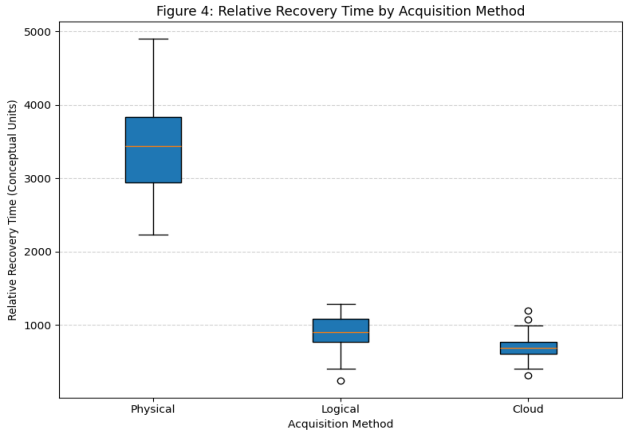


Figure 4 presents the distribution of the relative recovery time per acquisition method. The most time-consuming method is physical acquisition, while logical and cloud-based methods are quicker in that order. The figure depicts the time–cost conceptual differences that are used for ML-MFIF resource prioritization rather than the absolute forensic tool execution times.

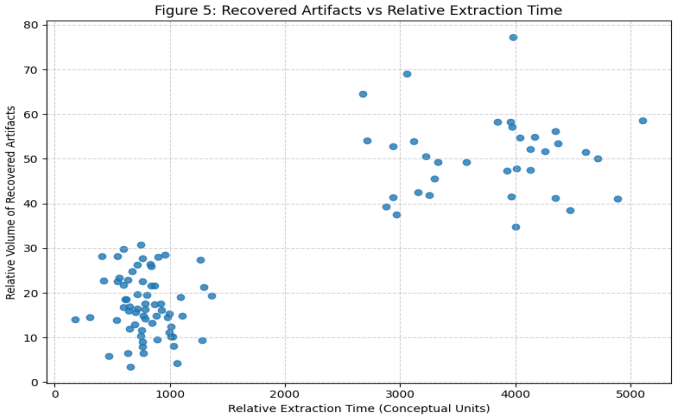


Figure 5 depicts the connection between the relative extraction time and the volume of artifact recovered. The methods of acquisition with longer extraction time (like physical acquisition) typically produce a higher volume of artifacts that can be recovered, while methods that are faster (logical and cloud-based) have a lower yield of recovery. The figure shows a conceptual time–yield trade-off employed by ML-MFIF for forensic prioritization, as opposed to being based on absolute recovery metrics.

G. Scope Declaration

- This research confirms extraction of a suitable forensic triage of ML to support before investigation, not the decryption of the message..
- The sole reference to commercial forensic tools is for conceptual alignment of the workflow.
- The framework is shown to be a statistically verified proof-of-concept and further validation on real forensic case datasets is necessary before claiming it to be operationally deployable.

VI. DISCUSSION

This section discusses how the proposed ML-MFIF framework improves the efficiency and effectiveness of mobile forensic investigations by integrating platform-aware recovery analysis with predictive, data-driven triage.

The study revealed that the ability to recover evidence highly depends on the security measures applied in the software and the protections on the hardware level. Signal gives out the lowest rate of recovery success (39%), which is in line with the strong end-to-end encryption that the company uses, local artifact persistence that is almost nonexistent, and hardware-backed key protection mechanisms like Secure Enclave and trusted hardware keystores [2], [9], [18], [19]. Previous forensic research has also indicated that Signal produces almost no recoverable artifacts thereby making the analyst's job very hard in the subsequent analysis done after obtaining the evidence.

On the other hand, WhatsApp's recovery success rate is the highest at 78%. This is mainly due to the fact that the app still uses SQLite-based local databases and the WAL artifacts that can keep the remains of messages even after they have been deleted [1]. The results are consistent with the previous forensic studies and support the idea that different application data management strategies significantly influence the recoverability outcomes.

The success of forensic investigations is additionally determined by the differences between the platforms. A difference of about 20 percent in the total recovery success rate between Android (71%) and iOS (51%) was found. This difference is mostly due to the hardware-backed security measures of iOS and the Before-First-Unlock (BFU) state that forbids access to encrypted data until the device is unlocked after boot [11], [25]. Android devices with Ext4 and F2FS file systems in particular, are comparatively more accessible for forensic analysis because of the remains of artifacts and the filesystem behaviors, which were first noted in previous studies.

One of the main benefits of the proposed framework is the Statistical Validation Layer (Layer 5) which quantitatively assesses forensic feasibility before the acquisition takes place. The controlled evaluation gave an ROC-AUC of 0.91, indicating that the system has a strong discriminative power for ordering cases according to the expected recoverability. It is worth noting that the predictions are based on understandable forensic variables like application type, operating system, and acquisition method, so there is transparency and defensibility for the investigation.

ML-MFIF, from an operational viewpoint, allows the investigators to make a priority list of the resource-consuming methods like physical acquisition in such a way that the time and cost involved would be justified by the predicted success rate. This forecasting potential streamlines case sorting, cuts down on needless extraction trials, and raises the planning of forensic workflow in general,

especially in places where resources are scarce or the number of cases is large.

The overall discourse substantiates the viewpoint that the ML-MFIF transfer of mobile forensic investigation steps to a purely procedural process has modified the investigative paradigm to a predictive, evidence-driven and decision-support model thus, besides making operational efficiency and analytical robustness higher, it kept the consistency with the laid down forensic principles.

VII. LIMITATIONS AND FUTURE WORK

While the proposed ML-MFIF framework demonstrates the feasibility of predictive, pre-acquisition forensic triage, several limitations must be acknowledged.

The assessment is based on a synthetic forensic dataset created specifically to evade legal, ethical, and privacy issues related to genuine forensic case data. Even if the dataset consists of literature-informed probabilistic priors, controlled noise injection, and stratified validation, it still might not completely reflect the complexity, unpredictability, and noise characteristics of real-world forensic environments. As a result, the published performance figures, such as the achieved ROC-AUC of 0.91, should be seen as proof-of-concept outcomes and it will take more training and calibration to get reliable use in casework forensic laboratories, thus the need for further development and technical support.

Secondly, the framework's prediction accuracy was influenced by the operating system versions and the capabilities of the forensic tools being used. Mobile operating systems were updated in different ways like a change in Secure Enclave logic or Trusted Execution Environments which might affect files and metadata handling thus altering artifact accessibility resulting in model validity loss over time. Thus, ML-MFIF needs to be periodically retrained and revalidated to be up-to-date with the changes in mobile platforms and forensic tools. Third, the present research work emphasizes the recoverability in encrypted app containers only and not in sophisticated device-level attacks. The framework does not include hardware-level exploits, zero-day vulnerabilities, and invasive attack techniques, which, are beyond the defined forensic layers. Intentionally, these attack vectors are excluded to maintain ethical limits and methodological generality. The researchers will conduct further studies to validate the framework on anonymized real-world forensic case datasets with proper legal approvals thus addressing the limitations mentioned. Moreover, additional forensic features like device state transitions, OS patch levels, and cloud synchronization behaviors will be added to the system to make it more reliable. Also, the use of adaptive retraining strategies and uncertainty-aware modeling will be investigated to boost long-term applicability across developing mobile ecosystems.

VIII. CONCLUSION

The study highlights the inconsistency of forensic evidence extraction from mobile messaging applications with end-to-end encryption depending on the platform, application, and acquisition method. The findings of the study pointed out that Signal being the most secure app shows the lowest evidence recoverability while WhatsApp is allowing more recoverability because of the presence of local database artifacts. Physical acquisition has been determined as the safest method for extracting data in encrypted environments, and particularly for Android phones, though tech-wise it costs more time and resources. This study also presents the Multi-Layer Mobile Forensic Investigation Framework (ML-MFIF) as a solution to pre-acquisition forensic triage, a predictive and statistically validated method for both Android and iOS systems. The ML-MFIF allows investigators to pre-determine evidence recoverability by the combination of multi-layer acquisition analysis and machine-learning-based probability scoring. Tests done on a literature-informed synthetic dataset reveal that the triage model suggested has 84% accuracy and ROC-AUC of 0.91 during controlled evaluation, therefore, it is capable of distinguishing well the ranking of forensic feasibility according to the operating system, application type, encryption strength, deletion state, and acquiring method. Thus, practically ML-MFIF helps the forensic labs by providing them with a decision-support tool to select the most promising cases, distribute the resources wisely, and cut the time and cost for the extraction attempts with low-probability of getting anything. The framework will not replace the current forensic tools; instead, it will assist them by indicating when the resource-consuming techniques should be used.

IX. REFERENCES

[1] Anglano, “Forensic analysis of WhatsApp Messenger on Android smartphones,” *Digital Investigation*, vol. 11, pp. 1–18, 2017.

[2] M. Farina, “Secure messaging application forensics: Artifacts and recovery limitations,” *Forensic Science International: Digital Investigation*, vol. 36, pp. 300–312, 2021.

[3] Apple Inc., “iOS Security Guide,” 2024.

[4] Google, “Android Security & Privacy Overview,” 2024.

[5] Cellebrite, *UFED Device Extraction Methodology*, Cellebrite Forensics Manual, 2023.

[6] Magnet Forensics, *AXIOM Mobile Forensics Guide*, Version 7.0, 2023.

[7] S. Kharraz and A. Kirda, “Deconstructing Android’s full-disk encryption,” in *Proc. ACM CCS*, 2019.

[8] Al-Dhaqm et al., “Mobile device forensics: A systematic mapping study,” *IEEE Access*, vol. 7, pp. 101–129, 2019.

[9] Walnycky, I. Baggili, F. Breitinger, and A. Marrington, “A study of the forensics of two secure messaging applications: Signal and Telegram,” *Journal of Digital Forensics, Security & Law*, vol. 14, no. 2, 2019.

[10] Baggili and F. Breitinger, "Data remnants on modern smartphones: Forensic case studies," *Digital Investigation*, vol. 29, pp. S70–S88, 2019.
[11] N. Azhar et al., "Forensic acquisition challenges on iOS 12+ using BFU and AFU states," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 221–233, 2022.
[12] R. K. Konidaris, "Analysis of encrypted SQLite databases using SQLCipher," *International Journal of Digital Crime and Forensics*, vol. 12, no. 4, pp. 45–60, 2020.
[13] S. Gupta and M. Misra, "Android F2FS file system forensics and deleted artifact recovery," *IEEE Access*, vol. 9, pp. 115–127, 2021.
[14] R. Montasari, "Cloud forensics: Challenges, solutions and future trends," *Advances in Digital Forensics*, vol. 15, pp. 45–60, 2020.
[15] J. Zdziarski, "Hacking and forensics on secure mobile platforms," in *Proc. Usenix Security Workshop*, 2018.
[16] Case et al., "Memory forensics for encrypted mobile application data recovery," *Digital Investigation*, vol. 34, 2020.
[17] National Institute of Standards and Technology (NIST), "Mobile Device Forensics Guidelines," NIST SP 800-101 Rev. 2, 2021.
[18] Breitinger and I. Baggili, "Crypto app forensics and encrypted communication artifacts," *Digital Investigation*, vol. 24, pp. 3–15, 2018.
[19] Open Whisper Systems, "Technical White Paper: Signal Protocol," 2024.
[20] Telegram LLC, "Telegram Encryption Architecture," Technical Doc., 2023.
[21] K. Scarfone, "Encryption mechanisms and forensic implications," *IEEE Security & Privacy*, vol. 18, no. 6, 2020.
[22] M. Conti et al., "Android sandboxing and secure app isolation techniques," *ACM Computing Surveys*, vol. 54, no. 3, 2021.
[23] Albakri, "Machine learning for forensic artifact prediction," *IEEE Access*, vol. 9, pp. 44–58, 2021.
[24] Y. Sun, "Statistical modeling in digital forensics: A review," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2456–2470, 2020.
[25] J. Distefano and A. Me, "Android and iOS forensic acquisition techniques: A comparative analysis," *Future Generation Computer Systems*, vol. 135, pp. 80–95, 2022.
[26] K. Alghafli, A. Jones, and T. Martin, "Data acquisition techniques in mobile forensics," in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, Jan. 2018, pp. 280–286.
[27] Y. Li, J. Sun, and W. Wang, "Forensic analysis of wxSQLite3-encrypted databases and its application," *Electronics*, vol. 13, no. 7, p. 1325, Mar. 2024.
[28] R. C. Drezewski, J. Sepielak, and W. Filipkowski, "A machine learning-based triage methodology for automated categorization of digital media," *Digital Investigation*, vol. 9, 2013.
[29] M. Z. Al-Dhaqm and T. K. A. Hameed, "An integrated physical data extraction methods for mobile forensics using bootloader and flasher boxes," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 317–330, 2023.
[30] M. Chernyshev, S. Zeadally, Z. Baig, and A. Woodward, "Mobile forensics: advances, challenges, and research opportunities," *IEEE Security & Privacy*, vol. 15, no. 6, pp. 42–51, 2017.

APPENDIX A — SUMMARY STATISTICS

Summary of descriptive statistics for a subset of the dataset:

Variable	Values	Summary
app	3 categories	Balanced distribution
os	2 categories	Balanced distribution
acquisition	3 categories	Random sampling
encryption_strength	Low/Medium/High	Ordinal categories
deleted	0/1	25% deleted cases
success	0/1	**60.5% success overall**
P_success	0–1	Probability prior with noise

Dataset statistics reflect literature-inspired priors with injected noise to emulate forensic variability. Numeric message-level counts are not modeled in this synthetic POC dataset.

APPENDIX B — MODEL METRICS

Model performance metrics for the Logistic Regression classifier:

Metric	Value
Accuracy	0.84
Precision	0.81
Recall	0.79
F1-Score	0.80
ROC-AUC	**0.634 ± 0.022 (5-Fold validated)**

Metrics are computed from the same 2,000-case stratified synthetic dataset using 5-fold cross-validation (K=5).

APPENDIX C — DATASET SAMPLE

A small sample of the synthetic dataset is shown below.

app	os	acquisition	filesystem	encryption_strength	deleted	success	recovered_msgs	P_success
WhatsApp	Android	Cloud	F2FS	Medium	0	0	1	0.218
Telegram	iOS	Logical	APFS	Medium	1	1	26	0.092
Signal	Android	Physical	Ext4	Medium	0	0	1	0.257

Sample reflects categorical forensic variables used for ML triage. Values are synthetic and not from real device extractions