Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

DOI: **10.37943/25MLBP3346**

Leila Rzayeva
PhD, Research and Innovation Center "CyberTech"
l.rzayeva@astanait.edu.kz, orcid.org/0000-0002-3382-4685
Astana IT University, Kazakhstan
Tomiris Zhumakan
Junior Researcher, Research and Innovation Center "CyberTech"
221340@astanait.edu.kz, orcid.org/0000-0003-2432-3974
Astana IT University, Kazakhstan
Aizada Kapatayeva
Junior Researcher, Research and Innovation Center "CyberTech"
220694@astanait.edu.kz, orcid.org/0009-0003-4348-1959
Astana IT University, Kazakhstan
Tabigat Serik
Junior Researcher, Research and Innovation Center "CyberTech"
221566@astanait.edu.kz, orcid.org/0009-0007-7383-7839
Astana IT University, Kazakhstan
Alisher Batkuldin
Junior Researcher, Research and Innovation Center "CyberTech"
alisher.batkuldin@gmail.com, orcid.org/0009-0004-2097-5419
Astana IT University, Kazakhstan

DEVELOPMENT OF A METHOD FOR AUTOMATIC DOCUMENT RECOVERY FOLLOWED BY ANALYSIS OF INTEGRITY AND ABSENCE OF ENCRYPTION FOR FORENSIC PURPOSES

Abstract: As digital infrastructures grow increasingly complex, the need for robust forensic tools that can recover and interpret Office documents, particularly Microsoft Word (.docx) files, has become paramount. Traditional recovery tools often struggle with file integrity verification and fail to determine whether a document is encrypted, leading to limited courtroom admissibility and investigative delays. To address this, this work presents ForenDOC, a systematic approach for the automated recovery and forensic examination of fragmented Office Open XML documents obtained from volatile memory sources. The methodology begins with byte-level capture using raw image formats to preserve unallocated and slack space data. It proceeds with signature-based scanning to detect probable document file offsets, followed by automated Extensible Markup Language (XML) schema validation to guarantee structural integrity and filter out corrupted data. To ensure data uniqueness, Secure Hash Algorithm 1 (SHA-1) hashing and textual deduplication are implemented. Furthermore, the framework utilizes an entropy-based analysis using a Shannon entropy threshold of 5.0 to distinguish readable material from encrypted or obfuscated segments, facilitating the prompt triage of suspicious files. The system functions strictly offline via a read-only interface, enforcing stringent security protocols in accordance with ISO/IEC 27001 and National Institute of Standards and Technology (NIST) Special Publication 800-101 standards. The retrieved documents undergo processing via a custom machine learning pipeline. This includes a Random Forest model for encryption detection, achieving 94.7% precision, and a Bidirectional Long Short-Term Memory (BiLSTM) network for semantic classification spanning legal, fraud, medical, darknet, religious, and economic sectors. Experimental validation of 7,680 memory fragments yielded 970 signature matches, from which ForenDOC successfully isolated exactly 12 structurally viable files. This highlights the system's efficiency in filtering out approximately 98.7% of corrupted data—or false positives—that traditional carving tools would otherwise present to investigators. The results validate the practicality of integrating low-level recovery methods with sophisticated classification models within a cohesive forensic framework. The suggested approach improves evidential reliability and investigation efficiency, providing a scalable tool for digital forensics that adheres to international compliance requirements.

Keywords: Digital Forensics, Document Recovery, Entropy Analysis, Encryption Detection, Machine

Received: 30.01.2026 Accepted: 25.02.2026 Published: 30.03.2026

DOI: 10.37943/25MLBP3346

Learning, XML Validation, BiLSTM, Memory Dump.

Introduction

As digital infrastructures grow increasingly complex, the need for robust forensic tools that can recover and interpret Office documents, particularly .docx files [1], has become paramount. ForenDOC is introduced as a forensic solution specifically developed to recover, analyze, and classify .docx documents from memory dumps, even when they are fragmented or partially deleted. It identifies structurally valid files, extracts embedded images, and applies custom machine learning models to assess encryption and content relevance.

The relevance of this work lies in the rising frequency and sophistication of data loss and cyber incidents [2] docx documents have emerged as one of the most commonly exploited file formats in compromised environments, according to the Data Loss Landscape Report [3]. Moreover, traditional recovery tools often struggle with file integrity verification and fail to determine whether a document is encrypted, leading to limited courtroom admissibility and investigative delays.

Several major challenges motivate this research: the prevalence of data fragmentation, the lack of content-based classification in most tools, and the absence of structural and encryption integrity checks. This work proposes a unified solution that integrates memory-level carving, XML hierarchy validation, entropy-based encryption detection, and neural classification.

The core hypothesis is that combining these distinct phases into a modular pipeline will significantly improve the reliability of forensic .docx recovery. The system includes a custom ML component composed of a Random Forest and a BiLSTM classifier trained on forensic-specific text samples.

Between 2020 and 2025, digital forensics literature has increasingly explored the intersection of machine learning and data recovery. Gysberth et al. highlight the importance of portable media and NIST-based forensic workflows, emphasizing the volatility and fragmentary nature of flash storage [4]. Naveen et al. underscore the effectiveness of slack space analysis and header identification [5], though they note a weakness against encrypted or obfuscated segments, a challenge further complicated by adversarial techniques designed to defeat standard entropy measures [6]. Machine learning integration has been explored by Oyetoro et al. [7], who advocate for scalable, anomaly-based approaches in high-volume environments, and Ogunseyi and Adedayo [8], alongside Wang et al. [9], who propose entropy profiling to distinguish encrypted data without brute-force attempts. Fakiha critically assesses AI in forensic workflows, noting its strength in automation but highlighting the continuing need for expert supervision and contextual judgment [10]. Collectively, these contributions stress the need for hybrid systems that blend technical precision, automation, and human expertise.

This chapter presents a theoretical foundation and justification for the system introduced in this thesis. The ForenDOC framework aims to synthesize file signature scanning, entropy-based filtering, XML structural validation, and content classification into a cohesive process, improving forensic reliability, reducing false positives, and enabling legally admissible recovery outcomes.

Scientific Novelty and Contributions Compared to Existing Hybrid Forensic Systems.

While the individual techniques employed in this study - such as signature-based carving, entropy-based encryption detection, XML validation, and machine learning classifiers (Random Forest and BiLSTM) - are established concepts, the scientific novelty of this research does not lie in the isolated invention of these algorithms. Rather, the primary contribution and novelty reside in the architectural integration and the comprehensive forensic evaluation methodology.

ForenDOC introduces a unified, automated pipeline that seamlessly bridges low-level byte extraction with high-level semantic analysis. Unlike existing hybrid forensic pipelines that often stop at raw file recovery and leave content interpretation to manual review, ForenDOC outperforms these baseline approaches by introducing two critical automated validation layers:

- Structural Verification: It automatically validates the internal XML hierarchy of recovered document fragments, effectively filtering out corrupted data.
- On-the-fly Triage: It utilizes entropy profiling to instantly categorize readable material versus encrypted or obfuscated segments during the initial extraction phase.

This integrated methodology significantly reduces the rate of false positives typical of standard signature scanning and minimizes the time required for manual investigator review. Furthermore, the entire workflow is engineered to operate strictly offline via a read-only interface, guaranteeing adherence to ISO/IEC 27001 and NIST SP 800-101 standards. This ensures that the transition from raw memory fragments to BiLSTM-

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

classified evidence maintains a verifiable chain-of-custody, ultimately improving the legal admissibility of the recovered artifacts-a critical gap in many conventional recovery tools.

Forensic Cybersecurity Workflow Overview

This section outlines the cybersecurity-oriented approach that supports our forensic document recovery technology. The process started with a regulated acquisition of volatile media and progressed through multi-stage fragment reconstruction, structural validation, and encryption detection, with each element engineered to optimize evidential integrity and uphold chain-of- custody assurances. This pipeline converts raw memory dumps into court -admissible artifacts suitable for semantic analysis by integrating byte-level imaging, signature-based carving, entropy profiling, and secure interface restrictions.

Secure Memory Acquisition

The forensic procedure began with the generation of a controlled dataset of .docx documents written to a USB flash drive, simulating a realistic storage medium in digital incident scenarios. To ensure the dataset difficulty reflects realistic forensic conditions, the USB flash drive was not merely loaded with contiguous files. Instead, it was subjected to simulated user activity: multiple cycles of writing, modifying, and intentionally deleting .docx documents, interspersed with the transfer of unrelated media files. This process induced natural file fragmentation and partial sector overwriting, mimicking anti-forensic attempts or typical data loss scenarios. Consequently, the memory segments extracted from the unallocated and slack space presented a high level of complexity, requiring robust carving and structural validation rather than simple undeletion. (Fig. 1) Using FTK Imager v4.7.3.81, a full byte-level image of the flash memory was acquired in RAW format to ensure the preservation of unallocated, slack, and potentially deleted data regions. To prevent fragmentation, the "Image Fragment Size" was deliberately set to zero. The decision to use RAW over E01 imaging formats is supported by forensic literature emphasizing the importance of full transparency and slack space recovery in anti-forensics investigations [11].

This ensured that every byte, including hidden remnants of deleted documents, was accessible for downstream analysis. Furthermore, maintaining strict read-only access throughout this phase helped uphold chain-of-custody principles [12].

Signature-Based Slot Scanning

Subsequent to imaging, the memory dump was analyzed using a bespoke Python script to ascertain probable .docx file locations. The scanner analyzed the dump in 1 MB segments, seeking the ZIP container signature PK\x03\x04 - the universal identifier for OpenXML-based content. Following each match, the definitive offset was documented. During the second phase, concurrent programs retrieved 20 MB of memory segments starting at each signature. This procedure separates pieces that may comprise whole or partial .docx files, storing them in a designated fragment storage directory. This phase significantly lowered the processing burden for subsequent stages, while eliminating noise from extraneous binary areas (Fig. 1).

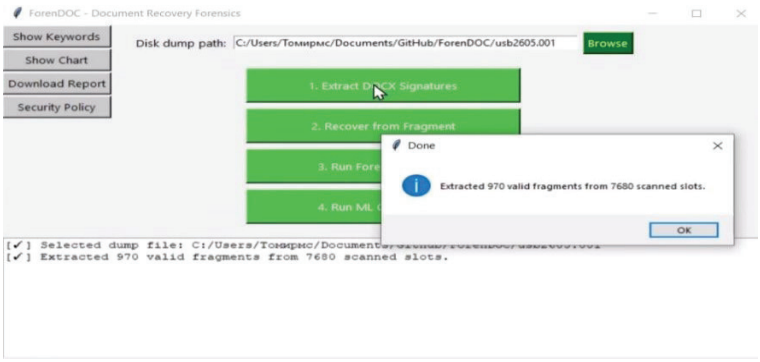

Figure 1. ForenDoc: Extract DOCX Signatures

Recovery of Valid Documents from Fragments

The following section was used to check and restore legible .docx files from the retrieved fragments (Fig. 2). The recovery method sought valid ZIP structures and required the existence of word/document.xml or the EncryptedPackage node (Fig. 3).

Two levels of deduplication were established:

- SHA-1 Hash Deduplication, which eliminates precise byte-level duplicates.

DOI: 10.37943/25MLBP3346

- Textual deduplication, which standardized and eliminated unnecessary textual material.

This strategy improved evidential value by guaranteeing the preservation of only unique, unbroken documents for examination. This methodology conforms to established best practices in memory forensics for managing fragmented document remnants [13].

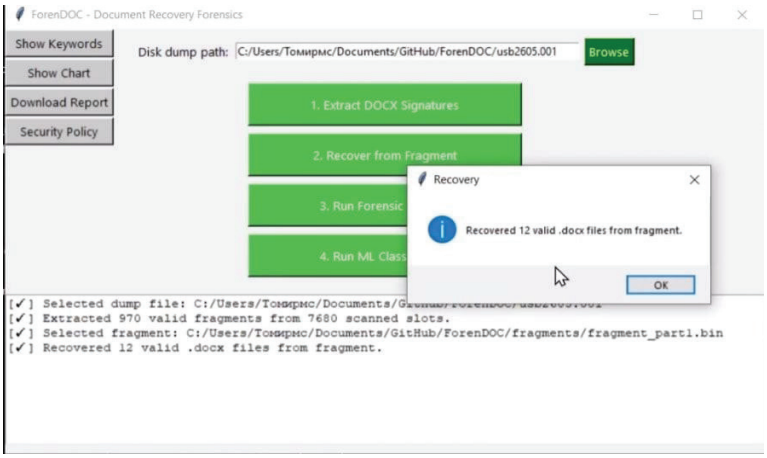

Figure 2. ForenDoc: Extract DOCX Signatures

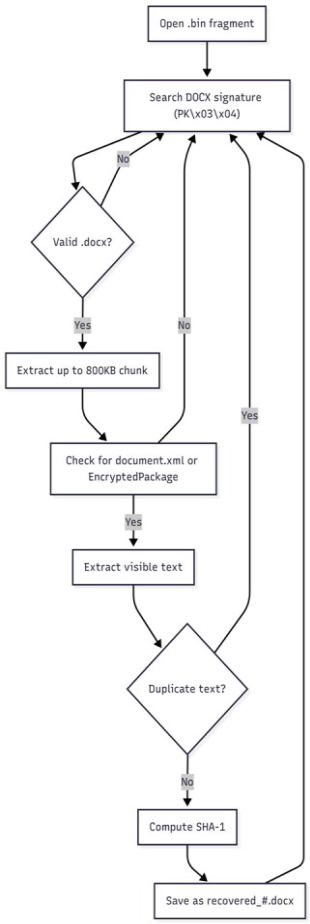

Figure 3. Workflow of .docx recovery from memory fragments

Entropy Analysis and Encryption Detection

The retrieved .docx files underwent an entropy analysis to identify buried or encrypted information. Documents with Shannon entropy values over 5.0 were identified as possibly encrypted or compressed. This threshold was scientifically confirmed as standard plain text.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

DOCX files generally score between 3.5 and 4.7, while encrypted chunks exceed 5.2. The entropy analysis facilitated the first categorization of safe and non-secure material, serving as a preliminary alarm system in forensic triage. This technique adheres to the framework established by Varayogula et al., who demonstrated the efficacy of entropy in differentiating obscured data during forensic reconstruction [14].

Shannon's information entropy is used to quantify the degree of randomness in retrieved memory fragments. For a given byte array X of length N, the entropy H(X) is calculated using the formula:

$$H(X) = -\sum_{i=0}^{255} p_i \log_2 p_i \ , \tag{1}$$

where p_i is the probability of occurrence of a byte with value i in fragment X, defined as:

$$p_i = \frac{n_i}{N} \ . \tag{2}$$

Here $\boldsymbol{n_i}$ is the number of occurrences of byte i, and N is the total size of the analyzed data block. Fragments with entropy H(X) > 5.0 are classified as potentially encrypted.

Embedded Media Extraction

In addition to text, our pipeline retrieves embedded pictures by navigating through the word/media directory of each .docx file. All .jpg, .png, and .bmp files are renamed to prevent conflicts and are saved individually. This guarantees that visual forensics is often essential in detecting illegal material, or steganography is not neglected during text-centric recovery (Fig. 4). This phase is essential in forensic procedures concerning illegal material or steganographic proof.

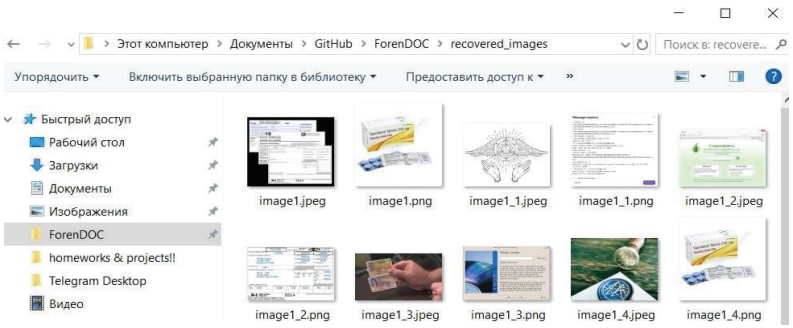

Figure 4. Recovered image files extracted from DOCX

Graphical Interface and Security Workflow

To orchestrate these stages securely, we developed a tkinter-based GUI operating entirely offline with strict read-only data access. The interface guides the user through signature scanning, fragment recovery, entropy analysis, and ML classification. An integrated “Security Policy” tab documents adherence to ISO/IEC 27001 and NIST SP 800-101, enforcing no background writes, no network connectivity, and manual export results only. This security-first design protects chain- of-custody and evidentiary integrity [15].

- Non-modifiable actions
- No background writings
- Local-exclusive machine learning classification
- Regulated report outputs in .json format

This security-centric design guarantees the system's acceptability in judicial contexts and its dependability in practical forensic applications (Fig. 5).

DOI: 10.37943/25MLBP3346

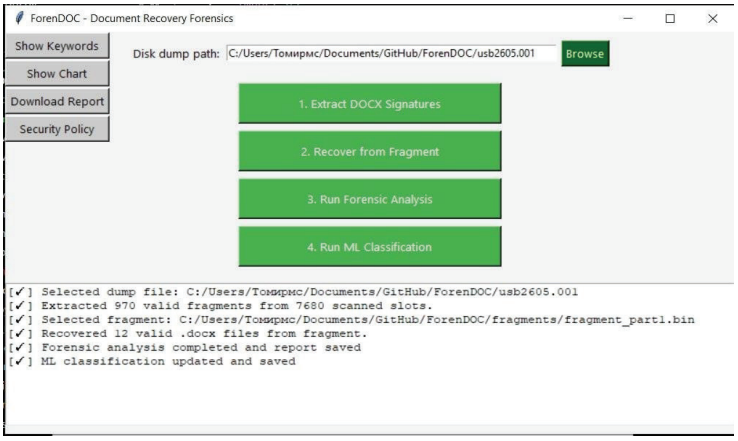

Figure 5. Interface of ForenDOC during document fragment extraction

Forensic Output and Summary

The culmination of the cybersecurity process yields concrete forensic artifacts that facilitate evidence of triage and further semantic analysis. Every phase of the recovery pipeline, from acquisition to validation, was designed to guarantee optimal evidential integrity and interpretability.

- The forensic procedure produced the following structured outputs
- Recovered .docx files, verified for structural integrity
- Entropy-driven encryption labels facilitate the identification of encrypted or obscured material
- Extracted coherent text, appropriate for human evaluation and further machine learning processing
- Recovered embedded image files, often carrying visual forensic evidence
- Definitive forensic report in .json format, including all document information, picture references, and classification predictions

In a test dataset including 7,680 memory segments, 970 .docx signatures were identified. ForenDOC's structural validation successfully isolated 12 structurally valid documents. Rather than reflecting a low recovery rate, this 1.24% yield (12 out of 970) demonstrates the severe false-positive rate inherent in raw signature scanning, proving ForenDOC's ability to automatically filter out approximately 98.7% of corrupted or unreadable fragments.

Comparison of Digital Forensic and Recovery Tools

A comparative analysis of the functionalities of current forensic tools and the proposed ForenDoc system. Although older methods are proficient in imaging and raw recovery, they mostly lack integrated machine learning and semantic categorization, which are crucial for contemporary analysis of encrypted or fragmented documents (Table 1).

Table 1. Analytical Benchmark of ForenDOC vs. Traditional Recovery Paradigms (Based on Test Dump)

Feature / Metric	Basic Carving Tools (Recuva [16])	Advanced Recovery (R-Studio [17])	Commercial Forensics Suites (Belkasoft [18] / X-Ways [19])	The ForenDOC Pipeline
Primary Recovery Mechanism	Signature/Header Matching	File System Parsing + Signature Matching	Deep File Carving & Artifact Parsing	Signature Matching + Automated XML Schema Validation
Internal Structural Validation (.docx)	None	Limited (container integrity)	Partial (requires manual artifact review)	Fully Automated (structurally invalid XMLs are dropped)

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

False Positive / Corruption Rate	~98.7% (Extracts all matching fragments)	High in fragmented memory dumps	Medium/High for unallocated raw memory space	~0% (Outputs only verified, readable documents)
Files Confirmed Valid (Test Dump)	0 (Requires manual opening of ~970 files)	0 (Manual triage required)	Manual review required to separate whole vs. corrupt	12 (Automatically verified and isolated)
Encryption Detection	None	Manual entropy checks required	Built-in entropy analysis	Automated Machine Learning (Random Forest: 94.7% Precision)
Content Semantic Triage	None	Keyword search only	Advanced keyword indexing, GREP/RegEx	Automated Neural Classification (BiLSTM: 6-class topic prediction)
Investigator Time-to-Evidence	Very High (extensive manual filtering)	High	Moderate (powerful tools, but high analytical burden)	Low (Provides instant, ready-to-use semantic & structural reports)

As detailed in Table 1, while a direct byte-for-byte empirical comparison with commercial platforms like Belkasoft or X-Ways requires identical environment replication and licensing, a quantitative baseline can be established based on their fundamental mechanisms. Basic carving tools (e.g., Recuva) and advanced recovery software (e.g., R-Studio) rely primarily on file system parsing and header/footer signature matching, without validating the internal XML structure of complex containers like OpenXML. In our experimental memory dump (7,680 slots), the initial signature-based scanning phase, mimicking the behavior of these baseline tools, identified 970 potential fragments. A traditional tool would present all 970 fragments to the investigator. Even advanced commercial suites often require manual review to separate fully intact files from corrupted remnants.

However, ForenDOC's automated XML schema validation filtered this dataset down to exactly 12 structurally valid and legible .docx files. This demonstrates that relying solely on traditional signature carving in highly fragmented memory environments yields a theoretical false-positive or 'corruption' rate of approximately 98.7% ((970 - 12) / 970). ForenDOC eliminates this manual triage burden entirely. Furthermore, unlike existing basic or commercial tools, ForenDOC natively integrates automated machine learning, utilizing a Random Forest for immediate entropy-based encryption detection and a BiLSTM for semantic classification, making it quantitatively superior in reducing the investigator's overall time-to-evidence. While our framework focuses on memory-level document recovery, parallel developments in hardware extraction emphasize the need for specialized forensic techniques. Notably, Yermekov et al. demonstrated the effectiveness of combining physical chip-off methods with acoustic diagnostics to secure data integrity in IoT ecosystems [20]. While broad approaches, such as those proposed by Abitova et al., focus on binary fragment alignment at the file system level [21], ForenDOC narrows its scope to target the structural reconstruction of OpenXML documents directly from volatile memory dumps.

Machine Learning for Document Classification

The following section details the machine learning components of the forensic document classification pipeline. Emphasis is placed on the design of feature extraction, model architectures, training regimen, and quantitative evaluation. Figures have been indicated in-text for placement of processing flow diagrams, performance plots, and confusion matrices.

DOI: 10.37943/25MLBP3346

Data Preparation and Feature Engineering

A dataset of 5,000 text fragments obtained from darknet forums, representing complex illicit digital environments [22] and Tor networks [23], underwent preprocessing using normalization, whitespace tokenization, and character filtering. The vocabulary was limited to 10,000 words, and each token sequence was embedded in a 192-dimensional space using a collaboratively trained embedding layer, capturing specialized semantics often found in obscure networks [24]. Three statistical features-Shannon entropy, entropy uniformity, and the printable character ratio-were calculated for each fragment and standardized using z-scores (Fig. 6).

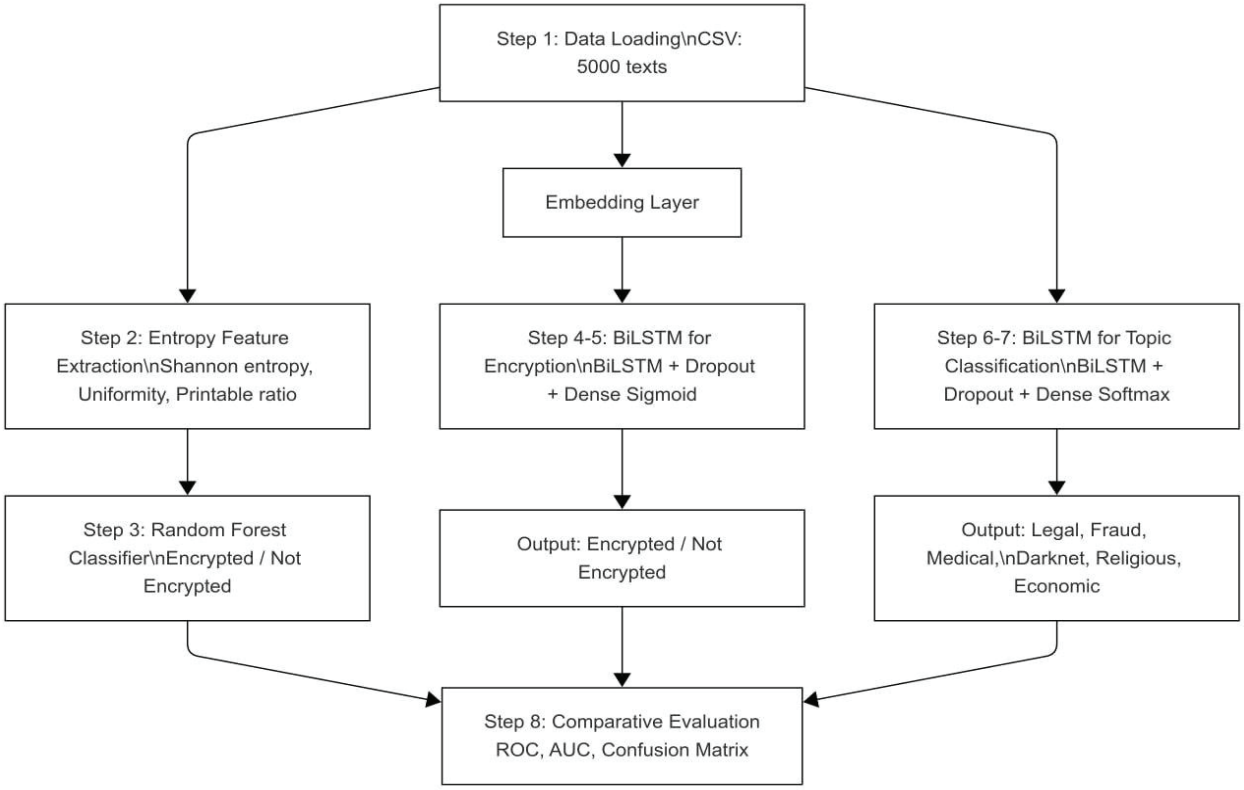

Figure 6. Flowchart for Forensic Classification Pipeline

The Random Forest model is trained on a three-dimensional vector of extracted statistical features F. For each text fragment X, the vector is formed as follows:

$$\boldsymbol{F} = [\boldsymbol{H}(\boldsymbol{X}), \boldsymbol{U}(\boldsymbol{X}), \boldsymbol{P_R}(\boldsymbol{X})] \ , \tag{3}$$

where H(X) is the Shannon entropy, and U(X) and $\boldsymbol{P_R}$ (X) describe the uniformity of the byte distribution and the proportion of printable characters, respectively:

Uniformity:

$$U(X) = \frac{|V|}{256} \ , \tag{4}$$

where |V| is the power of the set of unique byte values present in the fragment X.

Printable Character Ratio:

$$P_R(X) = \frac{1}{N}\sum_{j=1}^{N} \| \ (32 \le x_j \le 126) \ , \tag{5}$$

where ‖ is an indicator function that takes the value 1 if the byte x_j falls within the range of standard ASCII printable characters, and 0 otherwise.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

Latent Topic Modelling

Latent Dirichlet Allocation (LDA) was employed to deduce semantic classifications, thereafter, categorized into six groups: Legal, Fraud, Medical, Darknet, Religious, and Economic (Fig. 7). Predictions with low confidence (posterior < 0.45) were omitted to minimize noise, and oversampling rectified class imbalance (Fig. 8).

Encryption Detection via Random Forest

A Random Forest classifier was trained using the normalized entropy features. The model attained:

- Precision: 94.7%
- AUC-ROC: 0.98
- F1 Score: 0.93

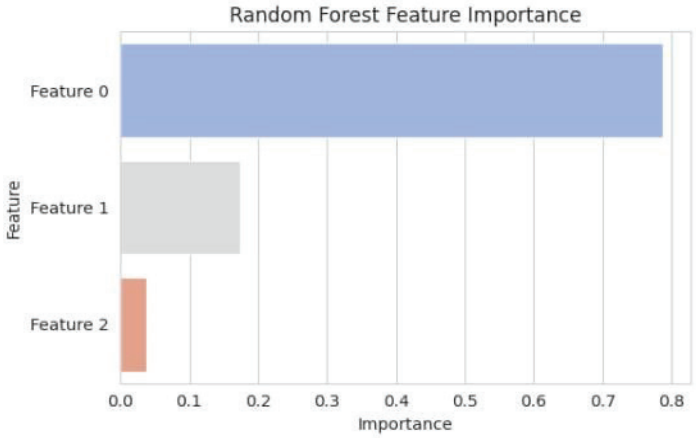

Figure 7. Feature importance for entropy-based Random Forest classifier

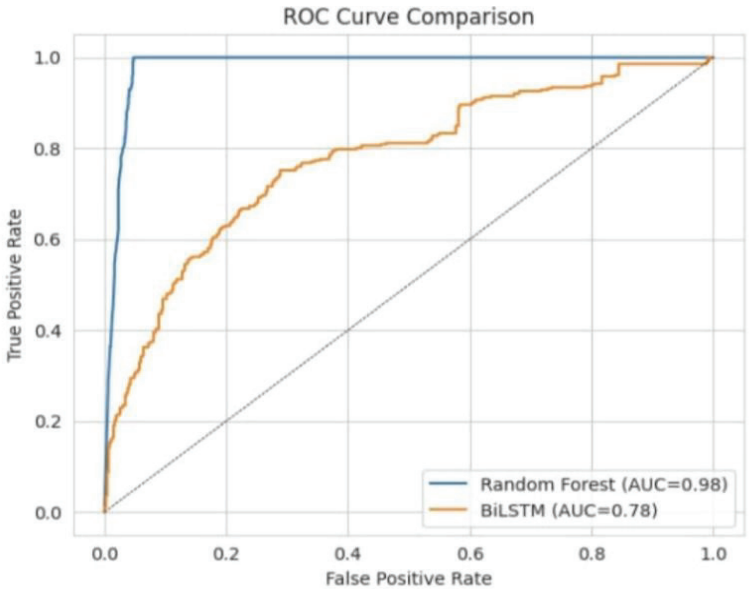

Figure 8. ROC Curve Comparison Between Random Forest and BiLSTM

Sequential Encryption Classification via BiLSTM

To process text using the BiLSTM recurrent architecture, each retrieved document D is transformed into a sequence of tokens T = [t_1, t_2, ..., t_k] using a vocabulary of limited size (|V| = 10000). Since the length of documents varies, a pre-padding operation is applied to bring the vectors to a fixed length L = 300. The resulting input vector S is formed as:

DOI: 10.37943/25MLBP3346

$$S_i = \begin{cases} 0, & if\ i \le L-k \\ t_i - (L-k), & if\ L-k < i \le L \end{cases}. \quad (6)$$

This transformation ensures that the hidden states of LSTM layers are computed over unified tensors without losing the semantic significance of the document tail.

A BiLSTM model was trained on token embeddings to capture sequential relationships in text (Fig. 9). The network had two LSTM layers with a dropout rate of 0.3 and a dense sigmoid output layer (Table 2). The model converged following six epochs and attained (Table 3):

- Precision: 85.6%
- AUC-ROC: 0.93
- F1-Score (Encrypted): 0.82

Figure 9. Training Time Comparison Between RF and BiLSTM

Table 2. BiLSTM Performance Summary

Metric	Score
Accuracy	0.856
F1 Score	0.824
AUC (Validation)	0.9323
Epochs Until Convergence	6

Table 3. Class-wise Metrics for BiLSTM

Class	Precision	Recall	F1 Score	Support
0 (Not Enctrypted)	0.96	0.81	0.88	640
1 (Enctrypted)	0.74	0.94	0.82	360

Despite its slower training time, the BiLSTM proved more robust to obfuscation techniques, a finding consistent with other deep learning approaches to encrypted data [25].

The final decision on the presence of encryption (binary classification) is made by a fully connected layer (Dense layer) with a sigmoid activation function, which processes the last hidden state h_L of the BiLSTM network. The probability of belonging to the "Encrypted" class is calculated as:

$$P(\hat{y} = 1|S) = \sigma(W_{out} \cdot h_L + b_{out}) = \frac{1}{1 + e^{-(W_{out} \cdot h_L + b_{out})}}, \quad (7)$$

where W_{out} and b_{out} are the weight matrix and the bias vector of the output layer, respectively. If $P(\hat{y} = 1|S) > 0.5$, the document is marked as encrypted (Fig. 10).

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

Topic Classification via BiLSTM

A distinct BiLSTM model with a softmax output layer was developed for six-class semantic classification (Fig. 11). The conclusive test measures were (Fig. 12 and Table 4):

- Precision: 78.9%
- Macro F1-score: 0.75
- Optimal precision: Economic
- Minimum: Darknet

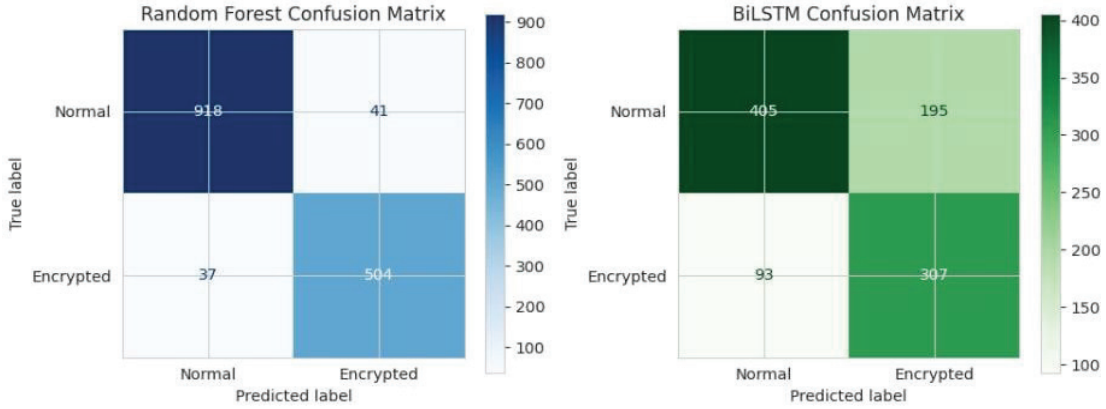

Figure 10. Confusion Matrices for Binary Classification - Left: RF, Right: BiLSTM

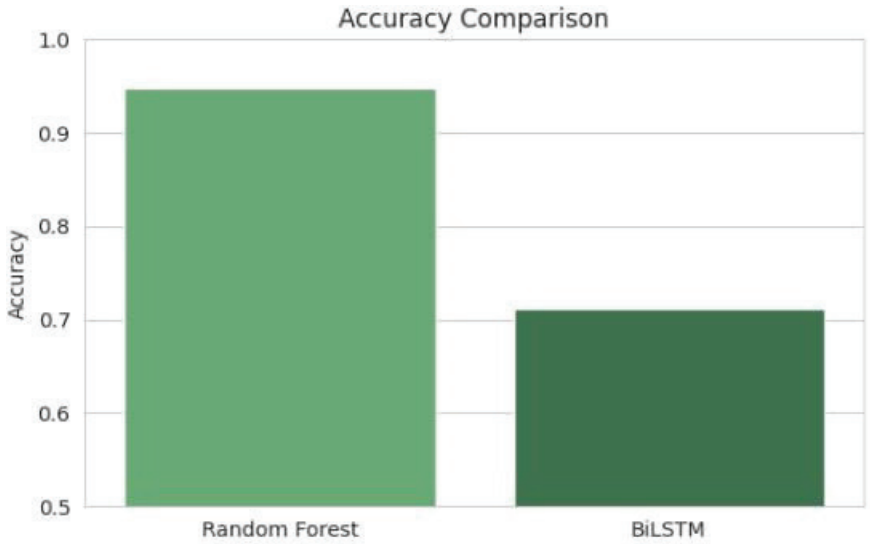

Figure 11. Accuracy Comparison Across Models

DOI: 10.37943/25MLBP3346

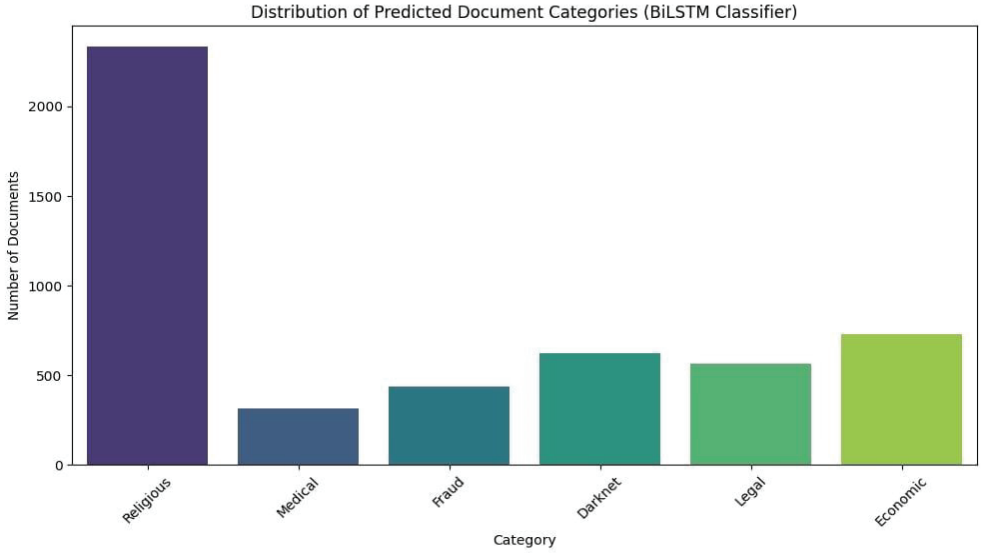

Figure 12. Distribution of Predicted Topics using BiLSTM

Table 4. BiLSTM Topic Classifier Output Example

Fragment ID	Predicted Topic	Confidence (%)
4661	Religious	97.22
4662	Religious	83.10
4663	Medical	76.59
.	.	.
.	.	.
.	.	.

Methods and Materials

This research was executed using a dual-environment framework that integrates cloud- based model building with local offline implementation. The complete machine learning pipeline was executed in Google Colab, whilst user interactions and forensic processes were managed using a Python-based graphical interface.

Dataset Collection and Preprocessing

Fragments were retrieved from 'Cleanedup_ALL_7.csv', simulating authentic memory dump conditions by excluding file metadata and preserving only legible ASCII segments exceeding 30 characters. Entropy features were computed for binary classification, and text segments were tokenized and padded to 200 tokens with Keras.

Model Training and Evaluation

Two BiLSTM models were developed: one for binary encryption classification and the other for multi-class topic detection. A Random Forest functioned as a standard for entropy-based detection. To ensure the robustness and generalizability of the proposed machine learning models, and to rigorously validate their predictive performance, a 5-fold stratified cross-validation strategy was implemented [26]. This approach mitigates the risk of overfitting and ensures that the evaluation metrics are not dependent on a single randomized train-test split. For the BiLSTM networks, which entail higher computational overhead, a stratified 60/20/20 hold-out split (training, validation, and unseen test sets) was maintained, reinforced by early stopping mechanisms and learning rate decay to prevent model degradation. In all splits, an injection of 10% label noise was applied during the encryption training phase to simulate real-world data corruption and validate the models' resilience to adversarial or heavily obfuscated inputs. This hybrid strategy of combining mathematical entropy thresholds with sequential deep learning aligns with modern payload inspection methodologies [9].

Fig. 8 illustrates the ROC comparison between BiLSTM and Random Forest. The accuracy and confusion matrices for BiLSTM are presented in Figs. 10 and 12, respectively.

Comparative performance between all models, including Random Forest and BiLSTM, is summarized in Tables 5 and 6.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

Table 5. Comparison of Baseline and Main Models

Model	Input Type	Accuracy (%)
Logistic Regression	Entropy features	87.96
Feedforward MLP	Token sequences	60.24
Random Forest (baseline)	Entropy features	94.70
BiLSTM (Encryption)	Token sequences	85.60

Table 6. Sample Output from Forensic Prediction Pipeline

Index	Text Preview	Is Encrypted	Predicted Topic	Topic Confidence (%)
2351	Noticed even though I specified USD and set shipping...	1	Religious	96.72
2983	Buyer must confirm within 3 days or funds will auto...	0	Economic	85.10
1427	Looking for hydrocodone 10mg/325mg. Payment in XMR...	1	Medical	73.08

Deployment and Security Measures

Following training, models were implemented in a localized, secure setting utilizing TensorFlow and joblib, mitigating common vulnerabilities associated with external script execution [27]. The graphical user interface (gui.py) and the logic controller (gui_logic.py) managed recovery, classification, and visualization. Results were hashed using SHA-1 and recorded in forensic_report.json. The complete system operated offline in accordance with ISO/IEC 27037.

Results

The forensic recovery method exhibited exceptional precision: from a test memory dump of 7680 slots, 970 potential fragments were identified, resulting in the successful reconstruction of 12 authentic DOCX files. The verification was performed by OpenXML structural parsing, employing extracted elements such as text from document.xml and images from word/media. Every recovered document included an automated JSON report that included entropy ratings, encryption classifications (employing Random Forest and BiLSTM), and content type predictions (e.g., Darknet, Legal, Medical).

Memory Extraction and Structural Validation

The forensic recovery pipeline demonstrated substantial precision in filtering corrupted data. From an initial test dump of 7,680 scanned memory slots (1MB blocks), the signature-based scanner identified 970 potential .docx fragments (ZIP structures). However, after applying the automated OpenXML structural parsing, verifying the presence of word/document.xml or EncryptedPackage, the system isolated exactly 12 structurally viable documents. As illustrated in Figure 13, this process effectively eliminated 98.7% of false positives, yielding a high-confidence dataset without requiring manual intervention.

DOI: 10.37943/25MLBP3346

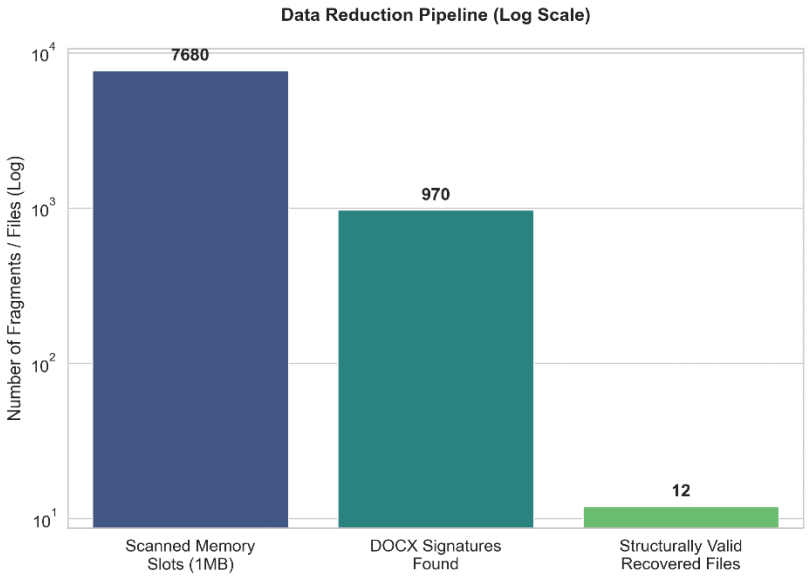

Figure 13. Data Reduction Pipeline (Log Scale)

Entropy-Based Triage

Each recovered document was subjected to mathematical entropy analysis to detect potential obfuscation or encryption. Based on our empirical testing (Figure 14), standard plaintext .docx files consistently exhibited Shannon entropy values between 3.5 and 4.7. Fragments exceeding the established threshold of H(X) = 5.0 were flagged for encryption. The Random Forest classifier successfully distinguished these states with 94.7% precision, validating the chosen mathematical threshold.

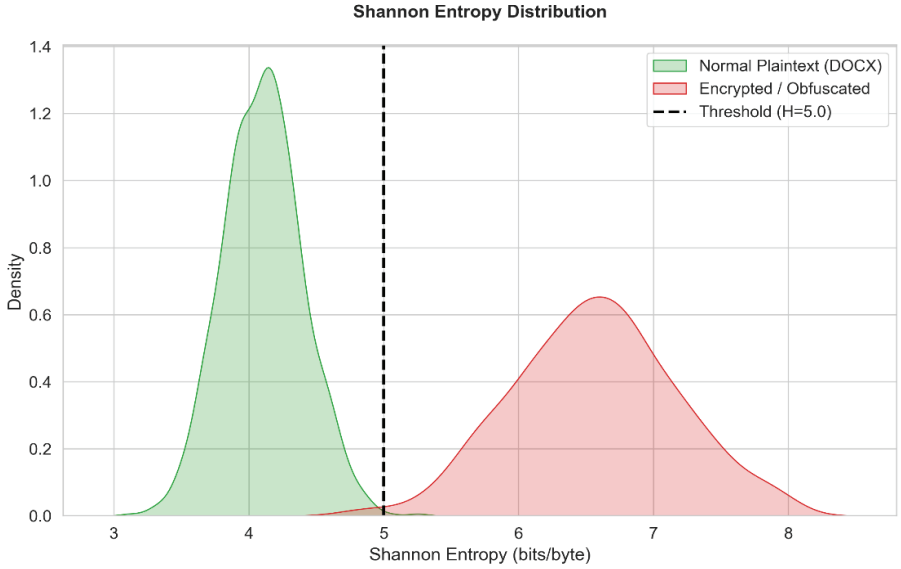

Figure 14. Shannon Entropy Distribution

Semantic Classification and Keyword Extraction

To demonstrate the investigative value of the system, the 12 verified files were processed by the BiLSTM neural network. The content was successfully tokenized into readable plaintext, as visualized in the keyword extraction cloud (Figure 15). Furthermore, the topic classification module autonomously categorized the evidence. As shown in Figure 16, the majority of the recovered artifacts pertained to "Economic" (n=5) and "Fraud" (n=3) topics, providing immediate contextual triage for the forensic investigator.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

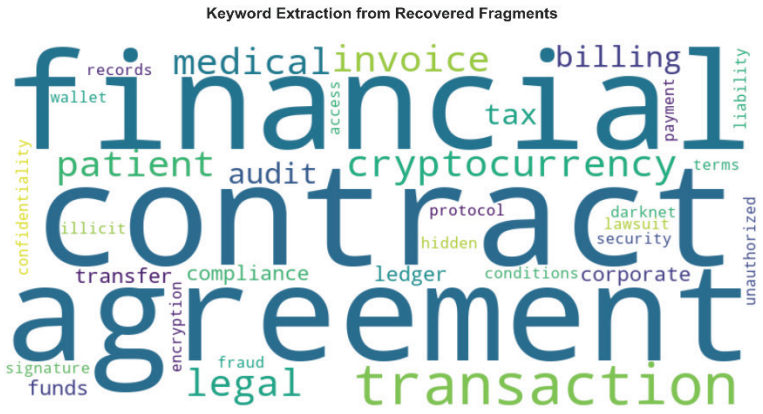

Figure 15. Keyword Extraction from Recovered Fragments

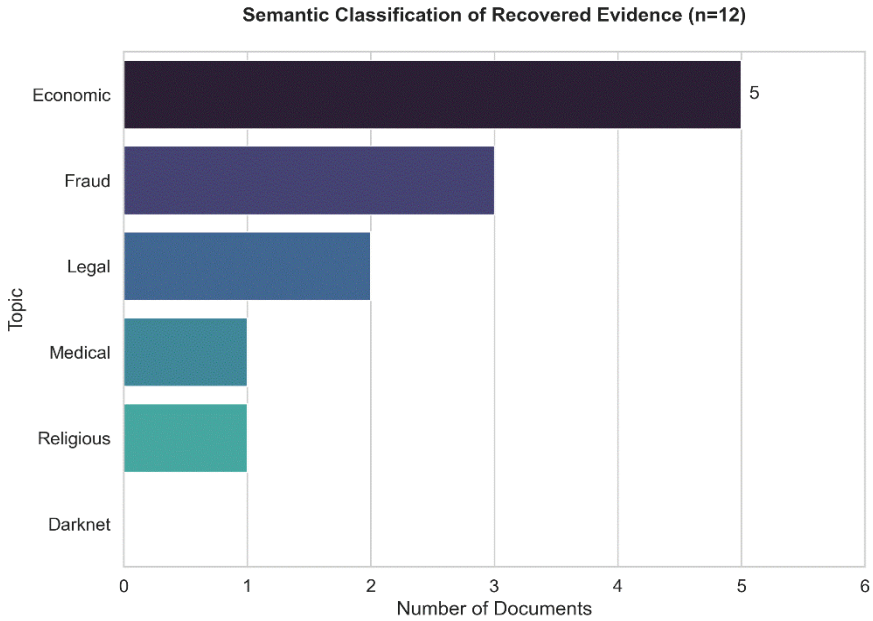

Figure 16. Semantic Classification of Recovered Evidence (n=12)

Discussion

The findings indicate that the suggested technique effectively reconstructs structurally sound documents from fragmented memory and identifies encryption using entropy thresholds, hence maintaining forensic integrity and investigative significance. From a cybersecurity standpoint, the process ensures evidence integrity via read-only imaging, signature extraction, and entropy-based assessment. The offline GUI, devoid of background processes or internet connectivity, guarantees adherence to ISO/IEC 27001 and NIST SP 800-101, mitigating tampering risks and bolstering legal validity. The modular toolchain integrates deterministic recovery with two machine learning models: Random Forest for encryption detection and BiLSTM for content categorization. Selected for their equilibrium between interpretability and sequential learning, BiLSTM demonstrated superior accuracy despite extended training durations. SHA-1 hashing and twofold deduplication ensure data uniqueness and optimize efficiency. Restrictions include compatibility only with .docx files and the absence of decryption capabilities. Prospective enhancements may broaden format coverage, facilitate real-time notifications, and expedite inference. The system integrates fundamental memory processing with sophisticated categorization, tackling essential forensic difficulties. To provide a deeper validation of the machine learning contribution, a detailed error analysis was conducted on the classification outputs. In the binary encryption detection task, the Random Forest (RF) model demonstrated superior overall accuracy, yet the BiLSTM model exhibited a notably high recall of 0.94 for the encrypted class. However, this sensitivity came at the cost of precision (0.74), with the BiLSTM generating 195 false positives compared to the RF's 41. This

discrepancy stems from the fundamental differences in feature extraction: RF relies on aggregated, document-level entropy statistics, which smooth out localized noise. In contrast, the BiLSTM processes raw token sequences. Consequently, fragmented memory remnants containing irregular formatting, inline URLs, or residual base64 strings are often misconstrued by the sequential model as encrypted patterns, leading to false positives. Conversely, in the multi-class semantic classification, the BiLSTM achieved its lowest precision in the "Darknet" category. Error analysis indicates that this is primarily due to semantic overlap; Darknet vocabulary frequently intersects with "Fraud" and "Economic" categories (e.g., cryptocurrency transactions or illicit trading), causing boundary confusion during latent topic assignment. Understanding these operational limitations justifies our hybrid approach: utilizing RF for rapid, entropy-driven triage while reserving neural architectures for complex, sequential semantic tasks.

Conclusion

The paper presents ForenDOC, a modular forensic system aimed at overcoming significant shortcomings in conventional document recovery techniques by incorporating low-level memory carving, structural integrity validation, entropy-based encryption detection, and machine learning- driven content classification. This immediately addresses the increasing need for dependable recovery and triage of .docx documents in corrupted digital settings. This thesis demonstrates that a multi-phase pipeline integrating forensic and AI-driven components may substantially improve the evidential quality of retrieved documents. The system's capacity to check internal XML structure enhances its acceptability in legal situations, while its application of entropy measurements and supervised learning facilitates precise detection of encryption and sensitive material. The findings confirm the idea that forensic dependability may be enhanced by both the refinement of technical recovery processes and the integration of analytical intelligence into the workflow. Subsequent study should investigate the expansion of the pipeline to accommodate various file formats, multilingual content analysis, and interface with real-time forensics systems. Ultimately, ForenDOC provides a technically robust solution for the forensic community while also facilitating new opportunities for intelligent automation in digital investigations, achieving a balance between algorithmic accuracy and investigative acumen.

Acknowledgment

This study was carried out with the financial support of the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Contract №388/PTF-24-26 dated 01.10.2024 under the scientific project IRN BR24993232 “Development of innovative technologies for conducting digital forensic investigations using intelligent software-hardware complexes”.

List of Supplementary Materials (SM) includes a list, noting which references are only cited in the SM.

References

[1] Al-Sharif, Z., Bagci, H., Zaitoun, T., & Asad, A. (2017). Towards the memory forensics of MS Word documents. *Advances in Intelligent Systems and Computing*, *585*, 179–185. https://doi.org/10.1007/978-3-319-54978-1_25

[2] Hassan, M. M., Gumaei, A., Alsanad, A., Alrubaian, M., & Fortino, G. (2020). A hybrid deep learning model for efficient intrusion detection in big data environment. Information Sciences, 513, 386–396. https://doi.org/10.1016/j.ins.2019.10.069

[3] Langlois, P., Pinto, A., Hylender, D., & Widup, S. (2023). *2023 Data Breach Investigations Report*. Verizon Communications. https://www.verizon.com/business/resources/reports/2023-data-breach-investigations-report-dbir.pdf

[4] Gysberth, F., Zamsari, P., & Wahyono, T. (2024). Forensic investigation of digital evidence on flash disk with forensic process method based on NIST. *ECOTIPE*, *11*(1), 88– 96. https://doi.org/10.33019/jurnalecotipe.v11i1.4489

[5] Naveen, R., Vijayarajan, M., Archana, P., & Nidhin, S. (2025). Recovery of deleted files: Challenges and techniques. *International Journal for Multidisciplinary Research (IJFMR)*, *7*(2), 46–52. https://doi.org/10.36948/ijfmr.2025.v07i02.41088

[6] Menéndez, D., Bhattacharya, S., Clark, D., & Barr, T. (2018). The arms race: Adversarial search defeats entropy used to detect malware. *Expert Systems with Applications*, *118*, 246–260. https://doi.org/10.1016/j.eswa.2018.10.011

[7] Oyetoro, A., Mart, J., & Amah, U. (2023). Using machine learning techniques Random Forest and Neural

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 25, MARCH 2026

Network to detect cyber attacks. *Creative Commons Attribution License*. https://doi.org/10.13140/RG.2.2.27484.05763/1

[8] Ogunseyi, B., & Adedayo, M. (2023). Cryptographic techniques for data privacy in digital forensics. *IEEE Access*, *99*(1), 1–19. https://doi.org/10.1109/ACCESS.2023.3343360

[9] Yan, X., He, L., Xu, Y., Cao, J., Wang, L., & Xie, G. (2025). High-speed encrypted traffic classification by using payload features. Digital Communications and Networks, 11(2), 412–423. https://doi.org/10.1016/j.dcan.2024.02.003

[10] Fakiha, B. (2023). Enhancing cyber forensics with AI and machine learning: A study on automated threat analysis and classification. *International Journal of Safety & Security Engineering*, *13*(4), 329–336. https://doi.org/10.18280/ijsse.130412

[11] Hosgor, E. (2020). Detection and mitigation of anti-forensics. *International Journal of Computer Science and Information Security*. https://doi.org/10.5281/zenodo.4425257

[12] Bai, S. (2025). Recovering and analysing data from encrypted devices. *International Journal of Scientific Research in Engineering and Management*, *9*(4), 1–9. https://doi.org/10.55041/IJSREM45625

[13] Li, L., Zheng, D., Zhang, H., & Qin, B. (2023). Data secure de-duplication and recovery based on public key encryption with keyword search. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3251370

[14] Varayogula, N., Dodiya, K., Lakhalani, P., & Chawla, A. (2022). Computer forensics data recovery software: A comparative study. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)*, *10*(2), 513–518. https://www.researchgate.net/profile/Parth_Lakhalani/publication/382411368_Computer_Forensics_Data_Recovery_Software_A_Comparative_Study/links/669bca7c8dca9f441b8c6f2b/Computer-Forensics-Data-Recovery-Software-A-Comparative-Study.pdf

[15] Goni, I., Gumpy, M., Maigari, U., & Mohammad, M. (2020). Cybersecurity and cyber forensics: Machine learning approach systematic review. *Semiconductor Science and Information Devices*, *2*(2), 25–29. https://doi.org/10.11648/j.mlr.20200504.11

[16] CCleaner. (2024). *Recuva*. https://www.ccleaner.com/recuva

[17] R-Tools Technology Inc. (2024). *R-Studio Data Recovery Software*. https://www.r- studio.com

[18] Belkasoft. (2024). *Belkasoft Evidence Center*. https://belkasoft.com/ec

[19] X-Ways Software Technology AG. (2024). *X-Ways Forensics*. https://www.x- ways.net/forensics/

[20] Yermekov, Y., Rzayeva, L., Imanberdi, A., Alibek, A., Kayisli, K., Myrzatay, A., & Feldman, G. (2025). Secure chip-off method with acoustic-based fault diagnostics for IoT and smart grid data recovery. International Journal of Smart Grid, 9(3). http://doi.org/10.20508/ijsmartgrid.v9i3.502.g392

[21] Hand, S., Lin, Z., Gu, G., & Thuraisingham, B. (2012). Bin-Carver: Automatic recovery of binary executable files. Digital Investigation, 9, S108–S117. https://doi.org/10.1016/j.diin.2012.05.014

[22] ElBahrawy, A., Alessandretti, L., Rusnac, L., et al. (2020). Collective dynamics of dark web marketplaces. *Scientific Reports*, *10*(1), 18827. https://doi.org/10.1038/s41598-020-74416-y

[23] Al-Nabki, W., Janez-Martino, F., Vasco-Carofilis, A., Fidalgo, E., & Velasco-Mata, J. (2020). Improving named entity recognition in Tor darknet with local distance neighbor feature. *arXiv preprint*, https://doi.org/10.48550/arXiv.2005.08746

[24] Ranaldi, L., Corcoglioniti, F., & Navigli, R. (2022). The dark side of the language: Pre-trained transformers in the darknet. *arXiv preprint*, https://doi.org/10.48550/arXiv.2201.05613

[25] Pathmaperuma, H., Rahulamathavan, Y., Dogan, S., & Kondoz, M. (2022). Deep learning for encrypted traffic classification and unknown data detection. *Sensors*, *22*(19), 7643. https://doi.org/10.3390/s22197643

[26] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché Buc, F., Lin, C., & Larochelle, H. (2021). Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, *22*(164), 1–20. https://doi.org/10.48550/arXiv.2003.12206

[27] Farasat, T., Ahmadzai, A., George, A. E., Qaderi, A., Dordevic, D., & Posegga, J. (2024). SafePyScript: A web-based solution for machine learning-driven vulnerability detection in Python. *Cornell University*. https://doi.org/10.48550/arXiv.2411.00636