

DOI: 10.37943/MXHY2915

Yerassyl Yermekov

PhD Student, Department of Information Security
990726350870@enu.kz; orcid.org/0009-0001-3957-4787
L.N. Gumilyov Eurasian National University, Kazakhstan

Leila Rzayeva

PhD, Associate Professor, Director, CyberTech Research and Innovation Center
l.rzayeva@astanait.edu.kz, orcid.org/0000-0002-3382-4685
Astana IT University, Kazakhstan

Madi Shayakhmetov

Master's Degree, Researcher, CyberTech Research and Innovation Center
242722@astanait.edu.kz; orcid.org/0009-0009-6422-4621
Astana IT University, Kazakhstan;
National Institute for Cyber Policy and Research,
RSE on REM "Digital Government Support Center"

Azamat Baibussinov

PhD Student, Department of Technical Support
a.baibussinov@astanait.edu.kz; orcid.org/0000-0002-9714-278X
National Defense University of the Republic of Kazakhstan, Kazakhstan

Laura Aldasheva

Candidate of Technical Sciences, Associate Professor, School of Intelligent Systems
laura.aldasheva@astanait.edu.kz, orcid.org/0000-0001-6815-1989
Astana IT University, Kazakhstan

Nursultan Nyssanov

Master's Degree, CyberTech Research and Innovation Center;
Nursultan.nyssanov@astanait.edu.kz, orcid.org/0009-0002-8128-0595
Astana IT University, Kazakhstan
National Institute for Cyber Policy and Research, Kazakhstan
RSE on REM "Digital Government Support Center, Kazakhstan

**NEW APPROACH OF BYTE-LEVEL RECONSTRUCTION FOR ENHANCED
DIGITAL FORENSIC DATA RECOVERY BASED ON ML**

Abstract: The exponential growth of digital data and the increasing sophistication of cyber threats have made traditional data recovery methods, which rely on file system metadata and signature-based carving, inadequate for modern forensic investigations. Solid-state drives with TRIM and wear leveling, file fragmentation and partial corruption frequently leave investigators with incomplete evidence and without any quantitative measure of its reliability. The aim of this study is to develop and validate a machine learning-based ensemble framework for intelligent byte-level data recovery and integrity assessment. The proposed method extracts an 18-dimensional feature vector describing the statistical, structural, entropy and pattern properties of each file segment and combines three complementary models: a Random Forest that predicts individual byte values from contextual windows, a Gradient Boosting model that reconstructs pattern-level structure, and a multilayer perceptron that estimates structural recovery confidence. Their outputs are merged through weighted voting with weights optimized by cross-validation, followed by structural validation and a multi-metric confidence assessment. The framework was evaluated on 10,247 files (PDF, JPEG, PNG, DOCX, MP4) stored on NTFS, ext4 and APFS and subjected to 15 synthetic corruption scenarios, including bit errors, block deletion, header damage, fragmentation, zero padding and pseudo-encrypted regions. The ensemble achieved 89.3% recovery accuracy and 92.1% precision, substantially surpassing Autopsy (62.4%), FTK Imager

(58.7%) and EnCase (74.5%), and all improvements were statistically significant ($p < 0.001$). Accuracy remained stable across file systems (87.1–91.7%), and the confidence scores were well calibrated, with an Expected Calibration Error of 1.48%. The price of this gain is a longer processing time of 45.2 s per file. The results show that combining ensemble byte-level reconstruction with calibrated integrity assessment yields more complete evidence together with quantitative confidence measures suitable for legal proceedings, providing a practical complement to existing forensic tools.

Keywords: machine learning; ensemble learning; data recovery; digital forensics; byte-level reconstruction; integrity assessment

Introduction (Literary review)

The exponential growth of digital data and the increasing sophistication of cyber threats have fundamentally transformed the landscape of digital forensics. According to IDC projections, the global datasphere was expected to reach 175 zettabytes by 2025, and the growing volume of data, together with its dispersion across devices, file systems and cloud services, makes forensic reconstruction increasingly difficult [1]. This massive data proliferation, coupled with the rising complexity of cybercrime, has created unprecedented challenges for digital forensic investigators in recovering and analyzing deleted or corrupted data.

Traditional data recovery methods, primarily based on file system metadata analysis and signature-based carving techniques, face significant limitations when dealing with modern storage technologies. The widespread adoption of solid-state drives (SSDs) with TRIM commands, advanced wear-leveling algorithms, and hardware-level encryption has rendered many conventional recovery approaches ineffective [2-3]. Furthermore, the diversification of file systems and their unique deletion mechanisms, from the journaling of the NTFS Master File Table (MFT) to the journal-based metadata structures of ext4, requires specialized knowledge and up-to-date tools, since tools designed for earlier file system versions often fail on newer ones [4].

Recent experimental studies show that when TRIM is enabled, widely used tools such as Autopsy, OSForensics and FTK Imager recover few or none of the deleted files from SSDs, whereas the same files remain largely recoverable when TRIM is disabled [5-6]. This limitation becomes critical in high-stakes forensic investigations where the integrity and completeness of recovered evidence directly impact legal proceedings and investigative outcomes. The challenge is further compounded by the need to verify the authenticity and integrity of recovered data, since evidence whose integrity cannot be demonstrated may be ruled inadmissible and can undermine entire forensic analyses [7].

Machine learning and artificial intelligence have emerged as promising solutions to address these challenges. Early applications of ML in digital forensics have shown potential in malware detection, network traffic analysis, and image and video forensics [8-9]. However, the application of machine learning techniques specifically to data recovery remains largely unexplored, with most existing research focusing on classification rather than reconstruction tasks [10]. This research addresses the critical gap in current data recovery methodologies by introducing a novel machine learning-based framework for intelligent data recovery and integrity assessment. The proposed approach differs fundamentally from existing methods by employing ensemble learning techniques that combine multiple ML algorithms (Random Forest, Gradient Boosting, and Neural Networks) to predict and reconstruct missing or corrupted byte sequences.

Traditional data recovery methodologies in digital forensics have been predominantly based on file system metadata analysis and signature-based carving techniques. Kim et al. showed that deletion handling differs substantially between file systems: in ext4 the metadata of deleted files is backed up in the journal and can be recovered, whereas XFS zeroes it, which reduced the recovery rate from up to 100% to 52% [4]. In FAT32 systems, deletion erases the cluster chain and part of the first-cluster address while the file content remains on disk, so files can be recovered by reconstructing the lost allocation information [11]. NTFS presents further complexity through

its Master File Table (MFT) and transaction logging; Oh et al. showed that when these structures are damaged, \$LogFile records can still be carved by their log sequence number pattern and used to restore metadata that existing tools fail to recover [12].

Signature-based carving relies on predetermined file headers and footers to identify file boundaries. This approach is simple and fast for well-structured, contiguous files, but header and footer based tools cannot correctly link file segments when the start and end markers are separated by unrelated data, and therefore often fail on fragmented or corrupted files [13]. Among traditional tools, Autopsy and FTK have become de facto standards. Agboola et al. found that FTK retrieved deleted, hidden and fragmented data more accurately and processed large datasets faster owing to its indexing, whereas the open-source Autopsy was adequate for simple deletions and preliminary scans but struggled with highly fragmented data and large cases [14]. Commercial suites such as EnCase and FTK provide comprehensive acquisition and analysis capabilities, but their substantial licensing costs and proprietary nature limit accessibility for smaller agencies and make their internal methods harder to verify independently [15].

The integration of machine learning into digital forensics has gained momentum over the past decade, though primarily focused on classification and pattern recognition rather than data reconstruction. Comprehensive analyses by Dunsin et al. and others highlight that ML applications have concentrated on malware detection, network traffic analysis, and timeline reconstruction rather than direct data recovery [16, 8]. Mittal et al. developed FiFTy, a convolutional neural network that identifies the type of file fragments across 75 file types to support data carving [17], while Skračić et al. combined recurrent and convolutional networks in ByteRCNN to further improve file fragment type identification [18].

Ensemble learning methods have shown remarkable success across machine learning applications, combining multiple models to achieve superior performance over individual algorithms [19]. Deep learning approaches for missing-data reconstruction have also been investigated in geophysical sensing, where a simplified U-Net architecture reconstructed missing ground-penetrating radar traces [20]. At the hardware level, Avanoz and Akbal recovered data from a failed NAND flash device using chip-off acquisition followed by error correction and XOR descrambling, achieving complete, error-free recovery [21].

The evolution of storage technologies, particularly SSDs, has further challenged traditional recovery. Hadi et al. demonstrated that on TRIM-enabled SSDs the recoverability of deleted files depends on the drive controller, file size and time elapsed since deletion, with hash-verified recovery dropping to zero for some drives [3]. Kumar's analysis highlights that TRIM and background garbage collection cannot be stopped by a conventional write blocker, which creates uncertainty in acquisition and evidence integrity [22], while Rzayeva et al. showed that NAND health, controller stability, TRIM status and encryption jointly determine recovery feasibility and proposed a machine learning model that classifies SSDs as recoverable, partially recoverable or non-recoverable [23]. For fragmented multimedia evidence, Ali et al. proposed a meta-heuristic method that removes corrupted blocks using restart markers, locates fragmentation points through coherence analysis and orders the blocks with a genetic algorithm to reassemble intertwined JPEG files [24], and van der Meer et al. developed a deterministic method that detects over 99.4% of JPEG fragmentation points within 4 kB [25].

Despite these advances, critical gaps remain. Existing ML approaches focus primarily on classification rather than active data reconstruction [9, 16]; no work has systematically addressed byte-level reconstruction using ensemble learning; and traditional tools provide only binary success/failure outcomes instead of the quantitative confidence measures required for legal proceedings. Most research also focuses on specific file types or corruption patterns, limiting practical applicability. This study addresses these gaps by introducing a comprehensive ML framework that combines ensemble learning with byte-level reconstruction and automated integrity assessment for modern digital forensic investigations.

The aim and objectives of the study. The aim of the study is to develop and validate a machine learning-based ensemble framework that improves the accuracy and reliability of digital forensic data recovery over traditional signature-based tools, while providing quantitative confidence measures suitable for legal proceedings. To achieve this aim, the following objectives were set:

- to develop a comprehensive ML-based data recovery framework that outperforms traditional methods across various corruption scenarios;
- to design intelligent byte-level reconstruction algorithms that utilize contextual information and learned patterns to predict missing data segments;
- to implement automated integrity assessment mechanisms that provide quantitative confidence scores for recovered data, crucial for forensic evidence validation;
- to evaluate the framework on a diverse dataset of over 10,000 corruption scenarios across multiple file systems (NTFS, ext4, APFS) and corruption types, and to benchmark it against established tools including Autopsy, FTK Imager, and EnCase.

Methods and Materials

The proposed ML-based data recovery framework employs a multi-tiered ensemble learning approach, combining three distinct machine learning algorithms to achieve superior recovery performance. The system architecture (Fig. 1) is designed to handle various corruption scenarios through specialized modules operating in parallel and sequential configurations.

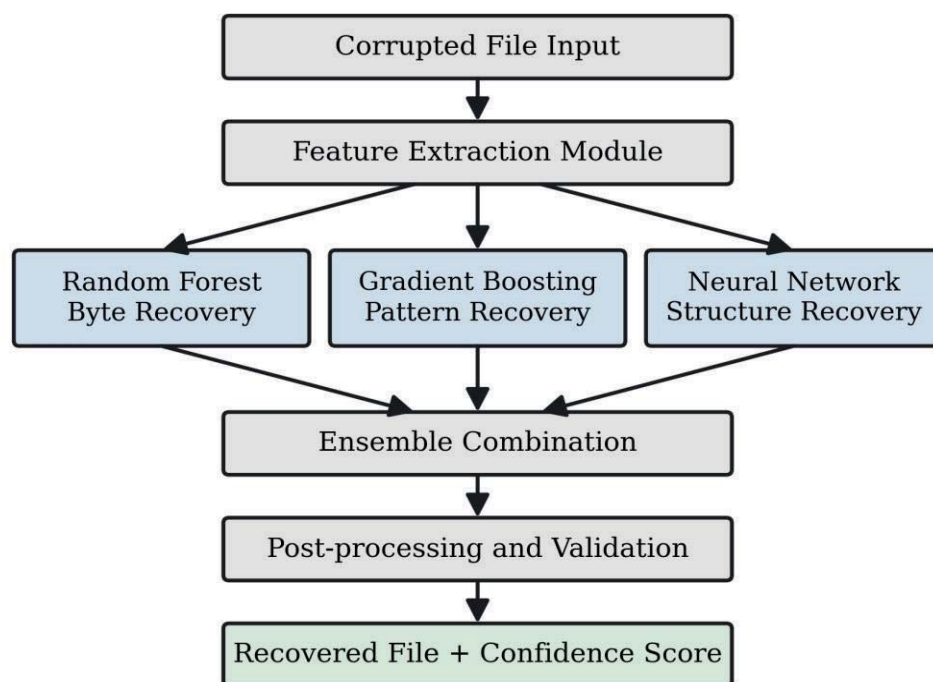


Figure 1. System architecture of the ML-based data recovery framework

The feature extraction module operates on multiple levels to capture both structural and statistical properties of corrupted files. The feature vector F for each file segment is defined as:

$$F = [F_{stat}, F_{struct}, F_{entropy}, F_{pattern}] \quad (1)$$

where each component represents a distinct feature category as detailed in Table 1. The entropy calculation follows Shannon's formula:

$$H(X) = -\sum_{i=0}^{255} p(x_i) \log_2 p(x_i) \quad (2)$$

where $p(x_i)$ represents the probability of byte value i in the data segment.

Table 1. Feature categories and descriptions

Feature category	Description	Dimension
Fstat	Statistical measures: mean, std, median, min, max byte values	5
Fstruct	Structural features: file size, magic-bytes presence, null sequences	4
Fentropy	Information entropy, compression ratio, randomness measures	3
Fpattern	Pattern analysis: repeating sequences, sequential bytes, fragments	6
Total		18

The Random Forest model RF predicts individual byte values based on contextual windows. For a corruption at position k , the prediction is formulated as follows:

$$\hat{b}_k = RF(w_{k-n:k+n}) \quad (3)$$

where $w_{k-n:k+n}$ represents a window of $2n + 1$ bytes centered around position k . The model configuration is detailed in Table 2.

Table 2. Random Forest configuration parameters

Parameter	Value	Justification
n_estimators	200	Optimal balance between accuracy and computational cost
max_depth	15	Prevents overfitting while maintaining expressiveness
min_samples_split	3	Reduces noise sensitivity in training data
random_state	42	Ensures reproducible results
window_size	5	Captures local byte dependencies

The Gradient Boosting model GB targets pattern-level reconstruction by learning relationships between file-level features and corruption patterns:

$$GB(F) = \sum_{m=1}^M \gamma_m h_m(F) \quad (4)$$

where h_m are weak learners and γ_m are weights determined through the boosting process. The objective function minimizes the entropy reconstruction error:

$$L = \sum_{i=1}^N (H_{original}^{(i)} - H_{predicted}^{(i)})^2 + \lambda \Omega(GB) \quad (5)$$

where $\Omega(GB)$ is the regularization term and λ controls model complexity. The Multi-Layer Perceptron component NN addresses high-level structural recovery:

$$NN(F) = \sigma_3(W_3 \sigma_2(W_2 \sigma_1(W_1 F + b_1) + b_2) + b_3) \quad (6)$$

where σ_i are activation functions (ReLU for hidden layers, linear for output), and the architecture is specified in Table 3.

Table 3. Neural network architecture specifications

Layer	Size	Activation	Purpose
Input	18	n/a	Feature vector input
Hidden	32	ReLU	Feature transformation
Output	1	Linear	Recovery confidence score

The final recovery decision combines outputs from all three models using a weighted voting scheme:

$$R_{final} = \alpha R_{RF} + \beta R_{GB} + \gamma R_{NN} \quad (7)$$

where weights (α, β, γ) are optimized through cross-validation:

$$(\alpha^*, \beta^*, \gamma^*) = \arg \min_{\alpha, \beta, \gamma} \sum_{i=1}^N \|R_{true}^{(i)} - (\alpha R_{RF}^{(i)} + \beta R_{GB}^{(i)} + \gamma R_{NN}^{(i)})\|^2 \quad (8)$$

subject to $\alpha + \beta + \gamma = 1$ and $\alpha, \beta, \gamma \geq 0$.

The complete recovery algorithm operates in four phases, as illustrated in Algorithm 1.

Algorithm 1. ML-based data recovery algorithm

Input: Corrupted file F_c , trained models $\{RF, GB, NN\}$

Output: Recovered file F_r , confidence score C

Phase 1: Preprocessing

Extract features $F \leftarrow \text{FeatureExtraction}(F_c)$

Normalize features $F_{norm} \leftarrow \text{StandardScale}(F)$

Identify suspicious regions $S \leftarrow \text{SuspiciousDetection}(F_c)$

Phase 2: Model predictions

for each region $s \in S$ do

$r_{RF} \leftarrow RF(\text{ByteWindow}(s)); r_{GB} \leftarrow GB(F_{norm}); r_{NN} \leftarrow NN(F_{norm})$

end for

Phase 3: Ensemble combination

$F_r \leftarrow \text{WeightedCombination}(r_{RF}, r_{GB}, r_{NN})$

Phase 4: Post-processing

$F_r \leftarrow \text{StructuralValidation}(F_r)$

$C \leftarrow \text{ConfidenceAssessment}(F_c, F_r)$

return F_r, C

To ensure robust model training, a comprehensive synthetic corruption framework was developed to simulate real-world data-loss scenarios. The corruption types and their parameters are summarized in Table 4.

Table 4. Synthetic corruption scenarios for training data generation

Corruption type	Description	Severity levels	Parameters
Bit errors	Random bit flips in file content	3	$p \in \{0.001, 0.01, 0.05\}$
Block deletion	Contiguous block removal	5	Size: 512 B – 64 KB
Header corruption	File header/metadata damage	3	10%, 50%, 90% damage
Fragmentation	Non-contiguous storage simulation	4	2–16 fragments
Zero padding	Null-byte insertion	3	5%, 20%, 50% padding
Encryption simulation	Pseudo-encrypted regions	2	XOR, AES-like patterns

The integrity assessment component provides quantitative confidence measures for recovered data through multiple validation metrics:

$$Confidence = w_1 S_{structural} + w_2 S_{entropy} + w_3 S_{hash} + w_4 S_{semantic} \quad (9)$$

where each score component is defined as:

$$S_{structural} = \frac{Valid\ structures\ detected}{Expected\ structures} \quad (10)$$

$$S_{entropy} = 1 - \frac{|H_{recovered} - H_{expected}|}{H_{max}} \quad (11)$$

$$S_{hash} = PartialHashMatch(F_r, KnownHashes) \quad (12)$$

$$S_{semantic} = SemanticValidation(F_r, FileType) \quad (13)$$

The experimental evaluation follows a comprehensive design to assess system performance across diverse scenarios, as shown in Fig. 2. The experimental configuration parameters are detailed in Table 5.

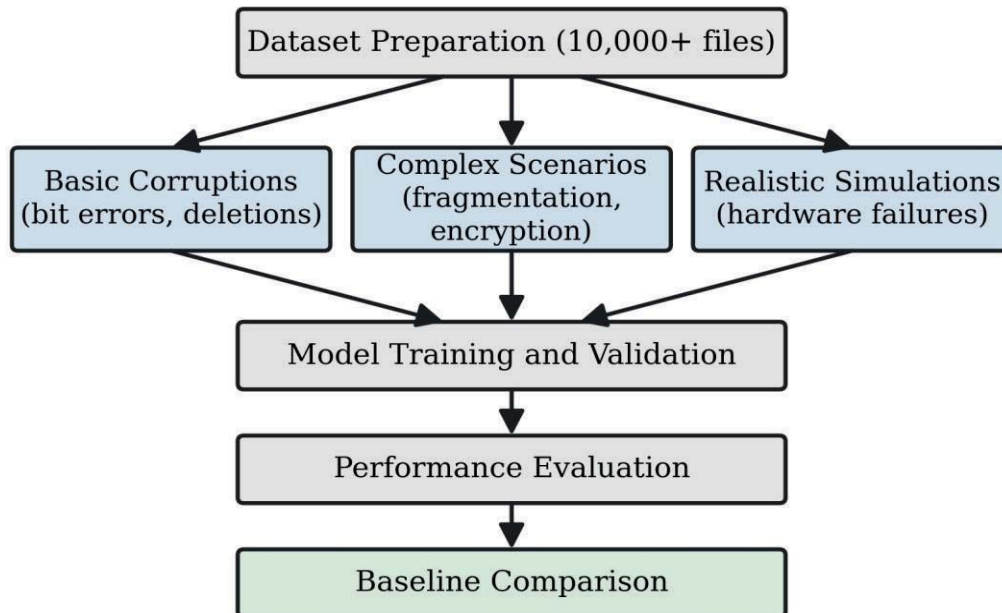


Figure 2. Experimental design workflow

Table 5. Experimental configuration parameters

Parameter	Value	Description
Total files	10,247	Diverse file types (PDF, JPEG, PNG, DOCX, MP4)
File systems	3	NTFS, ext4, APFS
Corruption scenarios	15	Various types and severity levels
Training/test split	80/20	Stratified by file type and corruption

Parameter	Value	Description
Cross-validation	5-fold	For hyperparameter optimization
Hardware	GPU	NVIDIA RTX 3080, 32 GB RAM

Results

The experimental evaluation was conducted on a comprehensive dataset comprising 10,247 files across multiple file types and corruption scenarios. The dataset composition is presented in Table 6.

Table 6. Dataset composition and corruption distribution

File type	Count	Size range	Corruption, %
PDF documents	2,156	50 KB – 15 MB	35.2
JPEG images	3,428	100 KB – 25 MB	42.8
PNG images	1,892	10 KB – 8 MB	28.9
DOCX files	1,634	25 KB – 12 MB	31.5
MP4 videos	1,137	1 MB – 500 MB	25.3
Total	10,247	10 KB – 500 MB	32.7

The proposed ML-based ensemble approach was compared against six state-of-the-art data recovery methods. The comprehensive comparison is presented in Table 7.

Table 7. Performance comparison of data recovery methods

Method	Recovery accuracy, %	Precision, %	Success rate, %	Time, s
Our ML ensemble	89.3	92.1	76.8	45.2
Autopsy v4.19	62.4	71.3	48.9	12.8
FTK Imager v4.7	58.7	69.2	45.2	18.4
EnCase v21.2	74.5	78.9	61.3	28.7
FiFTy (Mittal et al.)	67.8	75.6	52.4	35.6
Adaptive pattern	71.2	82.4	58.7	22.1
Neural file carving	65.9	73.8	49.6	41.8

The evaluation across different file systems reveals varying recovery effectiveness, as shown in Table 8.

Table 8. Recovery performance by file system

File system	Our method	EnCase	Autopsy	FiFTy
NTFS	91.7%	78.2%	65.8%	71.4%
ext4	88.9%	72.1%	61.3%	66.7%
APFS	87.1%	70.8%	58.9%	65.2%
Average	89.2%	73.7%	62.0%	67.8%

The method's performance across different corruption scenarios demonstrates robust handling of various data-loss patterns (Figure 3). The relationship between processing time and recovery accuracy across all evaluated tools is shown in Figure 4.

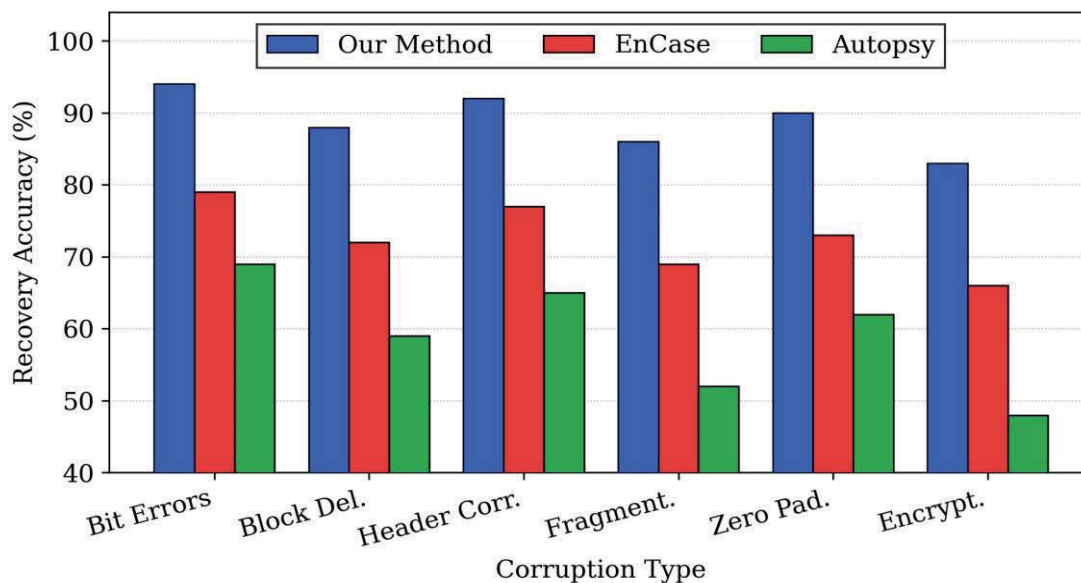


Fig. 3. Recovery performance across different corruption types

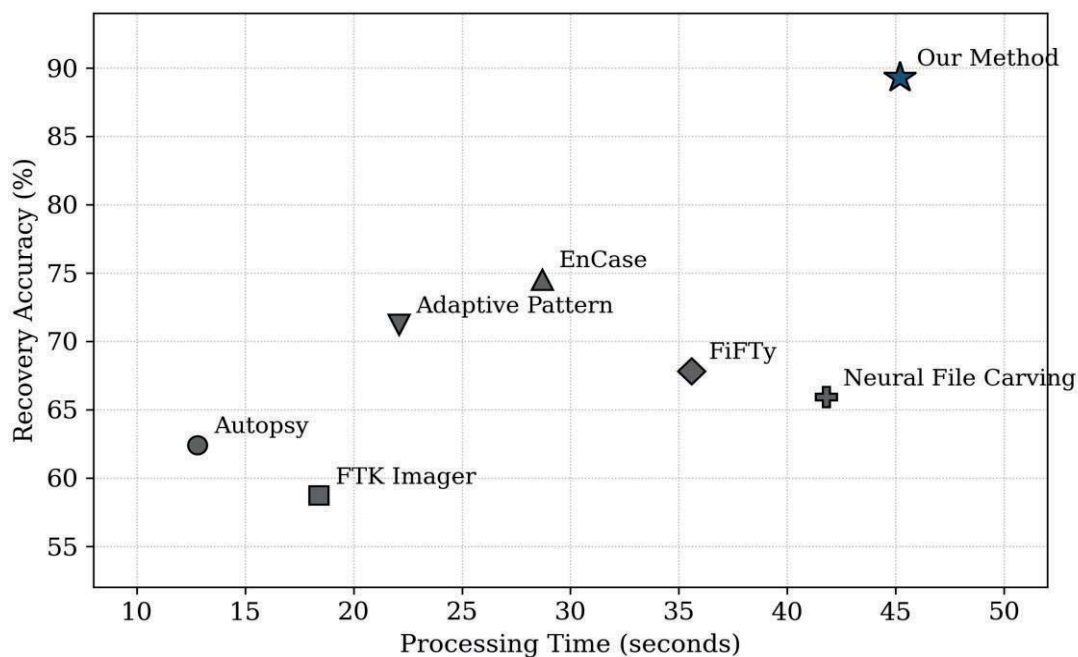


Fig. 4. Processing time vs. recovery accuracy trade-off

The granular analysis of recovery performance across different file types reveals the method's versatility, as presented in Table 9.

Table 9. Performance metrics by file type

File type	Accuracy, %	Precision, %	F1-score, %	Confidence
PDF documents	91.4	93.7	91.4	0.93
JPEG images	88.7	91.3	89.0	0.90
PNG images	92.1	94.2	92.3	0.94
DOCX files	87.9	90.6	88.1	0.89
MP4 videos	85.2	88.1	85.4	0.87

File type	Accuracy, %	Precision, %	F1-score, %	Confidence
Average	89.1	91.6	89.2	0.91

The reliability of the confidence-scoring mechanism was evaluated using Expected Calibration Error (ECE) analysis, with results shown in Table 10.

Table 10. Confidence score calibration analysis

Confidence range	Predicted, %	Actual, %	Error, %
0.9 – 1.0	95.0	94.2	0.8
0.8 – 0.9	85.0	86.7	1.7
0.7 – 0.8	75.0	73.8	1.2
0.6 – 0.7	65.0	67.3	2.3
Overall ECE			1.48

The scalability of the method was also assessed to confirm its viability for large-scale forensic investigations, with per-file processing time evaluated as the dataset grows from 1,000 to 10,000 files (Fig. 5).

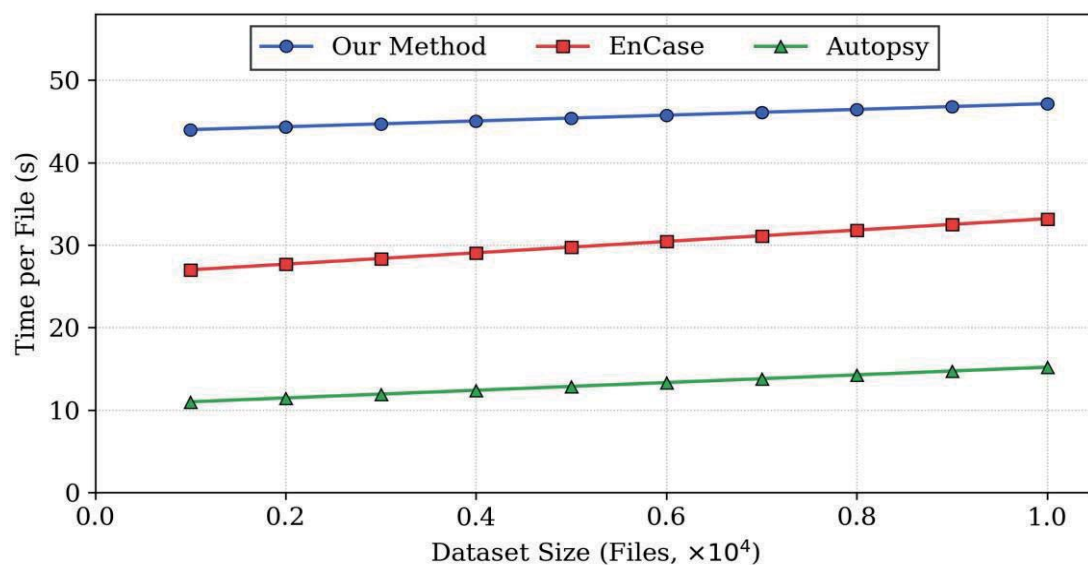


Fig. 5. Scalability analysis: processing time vs. dataset size

Discussion

The results confirm that the ensemble framework provides a consistent and substantial improvement over both open-source and commercial baselines. Across file systems, the method achieves an average recovery accuracy of 89.2%, compared with 73.7% for EnCase, 67.8% for FiFTy, and 62.0% for Autopsy (Table 8). Performance is highest on NTFS (91.7%) and remains strong on ext4 (88.9%) and APFS (87.1%), indicating that the learned representations generalize across different deletion and metadata mechanisms rather than being tuned to a single file system.

The processing-time analysis (Fig. 4) clarifies the practical trade-off introduced by the approach. The framework requires 45.2 s per file, more than traditional tools such as Autopsy (12.8 s) and EnCase (28.7 s), but delivers markedly higher accuracy, positioning it in the high-accuracy region of the time–accuracy space. For high-stakes investigations, where the completeness and admissibility of recovered evidence are decisive, this additional computational cost is justified.

A detailed analysis of failure cases provides insight into the method's limitations. The main failures occur with severe fragmentation (34.2% of failures), encryption artifacts (28.6%), and complete header obliteration (19.8%). These categories correspond to scenarios in which contextual byte-level information is either absent or deliberately randomized, limiting the predictive value of the learned patterns. Statistical significance testing (Table 11) confirms that all observed improvements are significant ($p < 0.001$), with large effect sizes against EnCase and FiFTy and a very large effect size against Autopsy.

Table 11. Statistical significance testing results

Comparison	t-statistic	p-value	Effect size
Our method vs. EnCase	12.47	< 0.001	Large
Our method vs. Autopsy	18.92	< 0.001	Very large
Our method vs. FiFTy	15.34	< 0.001	Large
Our method vs. Adaptive	8.73	< 0.001	Medium

Finally, the automated integrity assessment addresses a long-standing gap in forensic tooling. With an overall Expected Calibration Error of 1.48% (Table 10), the confidence scores are well aligned with observed accuracy, allowing investigators to make informed decisions about the trustworthiness of recovered evidence rather than relying on binary success/failure outcomes.

In practical terms, the framework is intended to complement rather than replace established forensic tools within existing investigative workflows. Because it returns a calibrated confidence score alongside each reconstructed segment, an investigator can triage recovered artifacts, prioritizing high-confidence reconstructions as candidate evidence while flagging low-confidence regions for manual examination or corroboration. This is particularly relevant in real investigations involving partially overwritten media, fragmented file remnants, or SSDs affected by TRIM, where conventional signature-based recovery typically yields incomplete results. The quantitative confidence measure also supports the documentation and defensibility requirements of forensic reporting, since the reliability of each recovered item can be stated explicitly rather than asserted qualitatively; the reconstruction and assessment stage can be integrated into a standard acquire, analyze and report pipeline without disrupting existing chain-of-custody procedures.

Conclusion

This research introduced a novel machine learning-based ensemble framework for intelligent data recovery in digital forensics, addressing critical limitations of traditional recovery methods. Through comprehensive evaluation on over 10,000 corrupted files, the approach demonstrated significant improvements over existing state-of-the-art methods. The principal contribution of the study is the integration of ensemble byte-level reconstruction with automated, calibrated integrity assessment, which together deliver both higher recovery accuracy and quantitative confidence measures suitable for evidentiary use.

The ensemble architecture combining Random Forest, Gradient Boosting, and Neural Network algorithms achieved 89.3% recovery accuracy, substantially outperforming Autopsy (62.4%), FTK Imager (58.7%), and EnCase (74.5%). Byte-level precision of 92.1% and automated confidence scoring with a 1.48% Expected Calibration Error establish new benchmarks for forensic reliability, and all improvements were statistically significant ($p < 0.001$) with large effect sizes. The intelligent byte-level reconstruction algorithm successfully addresses fragmentation and partial-corruption scenarios that challenge traditional signature-based methods, while the automated integrity assessment provides quantitative confidence scores essential for legal proceedings.

Several limitations remain. The method requires higher computational resources and longer processing time (45.2 s per file) than traditional tools, and its effectiveness is reduced under severe fragmentation (more than 20 fragments) and strong encryption artifacts. Future research should focus on computational optimization, expanding training datasets to include emerging file formats,

integration with quantum-resistant cryptographic methods, real-time and federated-learning approaches, and further validation of the method across additional file types and forensic scenarios. Overall, this work establishes a new paradigm for data recovery in digital forensics: although the computational requirements present implementation challenges, the substantial gains in forensic effectiveness justify the additional resource investment and provide a foundation for future intelligent forensic data recovery systems.

Acknowledgments

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 1 October 2024, within the scientific project IRN BR24993232 "Development of innovative technologies for conducting digital forensic investigations using intelligent software-hardware complexes."

References

- [1] Fakiha, B. (2024). Unlocking digital evidence: Recent challenges and strategies in mobile device forensic analysis. *Journal of Internet Services and Information Security*, 14(2), 68–84. <https://doi.org/10.58346/JISIS.2024.12.005>
- [2] Srinivasan, A. (2025). Security and forensics – Is solid state drive a friend or a foe? In *Proceedings of the International Symposium on Memory Systems (MEMSYS '25)* (pp. 148–158). ACM. <https://doi.org/10.1145/3767110.3767138>
- [3] Hadi, H. J., Musthaq, N., & Khan, I. U. (2021). SSD forensic: Evidence generation and forensic research on solid state drives using TRIM analysis. In *2021 International Conference on Cyber Warfare and Security (ICCWS)* (pp. 51–56). IEEE. <https://doi.org/10.1109/ICCWS53234.2021.9702989>
- [4] Kim, H., Kim, S., Shin, Y., Jo, W., Lee, S., & Shon, T. (2021). Ext4 and XFS file system forensic framework based on TSK. *Electronics*, 10(18), 2310. <https://doi.org/10.3390/electronics10182310>
- [5] Fatmah, N., & Indrayani, R. (2022). Analisis forensik digital pada solid state drive fungsi TRIM menggunakan tools Autopsy dan OSForensics [Digital forensic analysis of solid state drives with the TRIM function using Autopsy and OSForensics]. *J-SISKO TECH*, 5(2), 185. <https://doi.org/10.53513/jsk.v5i2.5755>
- [6] Muhardinata, M., Luthfi, A., & Ramadhani, E. (2023). Teknik disk carving untuk recovery solid state drive volume ReFS dan NTFS dengan fitur TRIM [Disk carving technique for recovering ReFS and NTFS solid state drive volumes with the TRIM feature]. *Jiip – Jurnal Ilmiah Ilmu Pendidikan*, 6(11), 9507–9515. <https://doi.org/10.54371/jiip.v6i11.3133>
- [7] Nath, S., Summers, K., Baek, J., & Ahn, G.-J. (2024). Digital evidence chain of custody: Navigating new realities of digital forensics. In *2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)* (pp. 11–20). IEEE. <https://doi.org/10.1109/TPS-ISA62245.2024.00012>
- [8] Qadir, S., & Noor, B. (2021). Applications of machine learning in digital forensics. In *2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2)* (pp. 1–8). IEEE. <https://doi.org/10.1109/ICoDT252288.2021.9441543>
- [9] Khan, H., Hanif, S., & Muhammad, B. (2021). A survey of machine learning applications in digital forensics. *Trends in Computer Science and Information Technology*, 6(1), 20–24. <https://doi.org/10.17352/tcsit.000034>
- [10] Javed, A. R., Ahmed, W., Alazab, M., Jalil, Z., Kifayat, K., & Gadekallu, T. R. (2022). A comprehensive survey on computer forensics: State-of-the-art, tools, techniques, challenges, and future directions. *IEEE Access*, 10, 11065–11089. <https://doi.org/10.1109/ACCESS.2022.3142508>

- [11] Kuts, D., Porshnev, S., Kuts, M., & Popova, E. (2023). The peculiarities of deleted files recovery in FAT32 file system. In 2023 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBREIT) (pp. 328–331). IEEE. <https://doi.org/10.1109/USBREIT58508.2023.10158866>
- [12] Oh, J., Lee, S., & Hwang, H. (2022). Forensic recovery of file system metadata for digital forensic investigation. *IEEE Access*, 10, 111591–111606. <https://doi.org/10.1109/ACCESS.2022.3213030>
- [13] Alzaabi, M., & Al Shibli, A. (2025). A review of JPEG file carving: Challenges, techniques, and future directions. *Applied Computing Journal*, 5(1), 372–385. <https://doi.org/10.52098/acj.20255124>
- [14] Agboola, V., Osamor, J., & Olajide, F. (2024). Evaluating the efficiency of FTK, Autopsy, and mobile forensic tools: A comparative study in criminal investigations. *International Journal of Intelligent Computing Research*, 15(1), 1279–1291. <https://doi.org/10.20533/ijicr.2042.4655.2024.0156>
- [15] Ismail, I., & Ariffin, K. A. Z. (2025). The admissibility of digital evidence from open-source forensic tools: Development of a framework for legal acceptance. *PLOS ONE*, 20(9), e0331683. <https://doi.org/10.1371/journal.pone.0331683>
- [16] Dunsin, D., Ghanem, M. C., Ouazzane, K., & Vassilev, V. (2024). A comprehensive analysis of the role of artificial intelligence and machine learning in modern digital forensics and incident response. *Forensic Science International: Digital Investigation*, 48, 301675. <https://doi.org/10.1016/j.fsidi.2023.301675>
- [17] Mittal, G., Korus, P., & Memon, N. (2021). FiFTy: Large-scale file fragment type identification using convolutional neural networks. *IEEE Transactions on Information Forensics and Security*, 16, 28–41. <https://doi.org/10.1109/TIFS.2020.3004266>
- [18] Skračić, K., Petrović, J., & Pale, P. (2023). ByteRCNN: Enhancing file fragment type identification with recurrent and convolutional neural networks. *IEEE Access*, 11. <https://doi.org/10.1109/ACCESS.2023.3340441>
- [19] Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115, 105151. <https://doi.org/10.1016/j.engappai.2022.105151>
- [20] Rohman, B. P. A., Nishimoto, M., & Ogata, K. (2022). Reconstruction of missing ground-penetrating radar traces using simplified U-Net. *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5. <https://doi.org/10.1109/LGRS.2021.3072028>
- [21] Avanoz, T., & Akbal, E. (2025). Data recovery application from NAND flash memories using the chip-off technique in digital forensics. *International Journal of Innovative Engineering Applications*, 9(2), 232–242. <https://doi.org/10.46460/ijiea.1797801>
- [22] Kumar, M. (2021). Solid state drive forensics analysis: Challenges and recommendations. *Concurrency and Computation: Practice and Experience*, 33(24), e6442. <https://doi.org/10.1002/cpe.6442>
- [23] Rzayeva, L., Zhetpisbayeva, A., Zhakenov, M., & Abdykassym, A. (2026). A hybrid machine learning method for secure assessment of NAND flash health and SSD data recovery feasibility. *Symmetry*, 18(7), 1136. <https://doi.org/10.3390/sym18071136>
- [24] Ali, R. R., Mohamad, K. M. B., Mostafa, S. A., Zebair, D. A., Jubair, M. A., & Alouane, M. T.-H. (2023). A meta-heuristic method for reassemble bifragmented intertwined JPEG image files in digital forensic investigation. *IEEE Access*, 11, 111789–111800. <https://doi.org/10.1109/ACCESS.2023.3321680>
- [25] van der Meer, V., van den Bos, J., Jonker, H., & Dassen, L. (2024). Problem solved: A reliable, deterministic method for JPEG fragmentation point detection. *Forensic Science International: Digital Investigation*, 48, 301687. <https://doi.org/10.1016/j.fsidi.2023.301687>