BIRZEIT UNIVERSITY

Faculty of Engineering & Technology
Computer Science Department

COMP438 - Digital Forensics

Final Project:

Explainable AI for File Type Classification in Digital Forensics Image Carving

Instructor: Dr. Mohammad Alkhanafseh

Prepared by:

Dena Abdo - 1192172
Motasem Ali - 1210341

Birzeit University
June - 2025

Abstract

In digital forensics, identifying file types from fragmented or partial data is a critical task—especially when metadata is missing or intentionally erased. Traditional file carving methods, which rely heavily on file headers and footers, often fail in such scenarios. While artificial intelligence (AI) has been introduced to enhance classification accuracy, many of these models operate as "black boxes," offering little insight into how decisions are made—an issue that limits their acceptance in legal contexts. This paper explores the use of **Explainable Artificial Intelligence (XAI)** to address this gap in transparency and trust. We focus on the **SIFT (Sifting File Types)** approach, which combines TermFrequency-Inverse Document Frequency (TF-IDF)-based feature selection with XAI techniques such as **LIME** and **SHAP** to highlight the most relevant features used in classifying file fragments. After identifying important byte patterns (features) from the file fragments, SIFT was used to train a model called a **Multilayer Perceptron (MLP)**. These features come from all types of files in the dataset — for example, some features might be common in JPGs, others in PDFs or DOC files. The MLP learns to recognize which **patterns of byte values** are linked to each file type. Then, when it sees a new fragment, it can guess which file type it belongs to. Since it can choose between many types (not just two), this is called **multiclass classification**. SIFT uses a standardized forensic dataset containing 47,482 fragments from 20 different file types, simulating real-world digital evidence scenarios. Our study begins by analyzing the limitations of traditional carving techniques such as Header/size-based, Hash based, Structure-based carving, then introduces the theoretical foundation of XAI and its relevance to digital forensics. We demonstrate how XAI enables interpretable and legally admissible AI outputs by explaining the reasoning behind each classification. Finally, the performance of SIFT was compared with other state-of-the-art approaches and the advantages in both accuracy and interpretability were discussed. This research shows that integrating explainability into forensic AI models not only improves transparency but also supports their admissibility and trustworthiness in legal investigations.

1. Introduction

Digital forensics plays a vital role in modern investigations by enabling the recovery and analysis of digital evidence from electronic devices. One of its key challenges is **file carving**—the process of reconstructing deleted or fragmented files from raw binary data, often in the absence of file system metadata [15]. This task becomes particularly difficult when files are fragmented or partially overwritten, where traditional signature-based methods (e.g., header/footer detection) fail to accurately detect or classify file types [2].

To address these limitations, researchers have introduced Artificial Intelligence (AI) and Machine Learning (ML) techniques to improve the automation and accuracy of fragment classification [2], [5]. However, many of these models act as **black boxes**, offering no insight into the reasoning behind their predictions. In digital forensics, where **transparency, reproducibility, and legal admissibility** are crucial, this lack of interpretability presents a serious concern for both investigators and courts [2], [4].

Explainable AI (XAI) has emerged as a solution to this trust gap. By providing interpretable justifications for AI predictions using techniques such as **LIME** (Local Interpretable Model-Agnostic Explanations) and **SHAP** (SHapley Additive exPlanations), XAI makes it possible to align AI outcomes with the evidence standards required in legal contexts [2], [17]. This is especially relevant in high-stakes forensic tasks where decisions must be investigated in court.

In this research, we explore the potential of **SIFT (Sifting File Types)**—an XAI-based approach for file fragment classification that combines TermFrequency-Inverse Document Frequency (**TF-IDF)** feature extraction, XAI interpretability tools, and a **Multilayer Perceptron (MLP)** model. SIFT is designed to not only classify file fragments with high accuracy but also **provide interpretable insights into why a fragment is classified a certain way**, addressing both performance and legal transparency.

The objectives of this paper are to:

- Evaluate current AI/ML capabilities in file type classification from fragmented data;
- Identify the limitations of traditional and black-box AI methods;
- Propose and explain a transparent classification pipeline based on XAI principles;
- Analyze its forensic and legal significance in digital investigations.

Ultimately, our goal is to demonstrate that explainable models like SIFT can improve not only classification performance but also the **legal defensibility and trustworthiness** of digital forensic analysis [4], [17].

2. Literature Review

2.1 AI and ML in Digital Forensics

AI and ML have emerged as transformative tools in DF, especially for **handling large datasets**, **identifying patterns**, and **automating investigative tasks**. Research highlights their effectiveness in tasks such as anomaly detection, image and text classification, behavioral analysis, and prioritization of forensic evidence [1][2][5]. Models like convolutional neural networks (**CNN**s) and support vector machines (**SVM**s) have shown success in multimedia analysis, including face and object detection [2].

2.2 Traditional File Carving Methods

Traditional file carving relies on identifying known file headers and footers (e.g., JPEG's \xffd8 to \xffd9). These methods work well on intact files but struggle with **fragmented files**, **missing footers**, or **unknown formats**. Additionally, they are often time-consuming and require expert manual intervention, leading to scalability issues [2].

2.3 Use of AI in File Carving and Fragment Classification

Recent studies have applied supervised ML models for file fragment classification, showing improved accuracy over signature-based methods. CNNs and deep learning approaches can learn complex patterns in binary data, enabling classification of fragment types even in the absence of clear signatures. **However, many such models lack transparency, limiting their acceptance in forensic practice** [1][5].

3. Background

3.1 Overview of file fragmentation and carving techniques

File carving is a key process in digital forensics used to recover files without relying on file system metadata. It works by analyzing raw data blocks and identifying file structures based on signatures, content, or statistical patterns. Several carving methods have evolved based on the level of fragmentation and data integrity.

- **Signature-based carving** searches for known header and footer patterns to extract files, assuming they are stored contiguously. Tools like *Scalpel* and *Foremost* implement this approach.
- **Header/size-based carving** estimates the file size from a known header and extracts the expected data range, though it assumes minimal fragmentation.
- **Hash-based carving**, used in cases like the 2006 DFRWS challenge, compares hashes of known files to disk fragments to identify matches—even partial ones. For example:
    1. JPEG—"xFFxD8" header and "xFFxD9" footer
    2. GIF—"x47x49x46x38x37x61" header and "x00x3B"
    3. PST—"!BDN" header and no footer
    4. If the file format has no footer, a maximum file size is used in the carving program
- **Structure-based and semantic carving** analyze the internal structure or language within a file to distinguish fragments, improving precision and reducing false positives.
- **Content-based carving** examines the characteristics of each cluster (e.g., size, character frequency) to group and reassemble them, useful for recovering fragmented files.

- **Graph-theoretic methods** apply algorithms like shortest path first or Hamiltonian path (if the path starts and ends at different vertices, it's a Hamiltonian path) to determine the most likely order of fragments, particularly for images and text.
- **Bifragment gap carving** focuses on recovering files with known headers and footers that exist in two fragments close to each other.

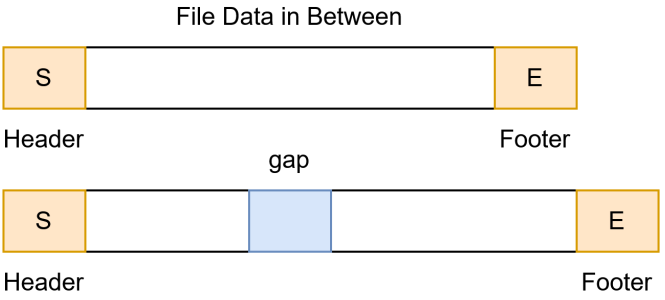


Figure 1: *Bifragmented file.*

- **Smart carving** combines multiple methods to process both fragmented and non-fragmented files, using preprocessing, classification, and reassembly phases.

These techniques offer varying effectiveness depending on the file system, device type, and level of fragmentation. Their growing sophistication supports the recovery of deleted or hidden data crucial for forensic investigations. [6]

3.2 Why image file type classification matters (e.g., child exploitation, malware)

In digital forensic investigations, classifying file types—especially image files—plays a critical role when analyzing raw data. Often, forensic analysts work with disk images that contain vast amounts of unstructured binary data, such as fragments from deleted or unallocated space, where file system metadata is unavailable. Identifying the type of each fragment helps prioritize relevant content for examination [7]. This becomes even more vital when malicious actors intentionally rename or hide files (e.g., executables posing as image files) to evade detection, making content-based classification more reliable than file extensions alone [8]. In sensitive cases like child exploitation, investigators deal with thousands of image fragments, making it impractical to examine all files manually. AI-based image classification enables faster sorting, helping law enforcement focus on fragments most likely to contain evidence [9]. Moreover, using explainable AI models to support classification adds legal weight to forensic results, as courts increasingly demand transparency and scientifically valid reasoning behind presented evidence [10].

3.3 Explainable AI (XAI) vs Interpretable vs Black Box

The adoption of AI in legal proceedings depends on model transparency. Black-box models are often unsuitable due to their lack of interpretability. Explainable AI (XAI) addresses this by providing **human-understandable justifications for model outputs**. Methods include SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), and rule-based models that prioritize clarity over complexity. In DF, where evidence must be clear and acceptable in court, interpretability is vital for ensuring admissibility and credibility. [3][4].

A black box model in AI refers to models that utilize complex algorithms, such as neural networks. These algorithms typically yield high accuracy; however, the rationale behind their decisions remains ambiguous. In contrast, white box models allow for interpretability, highlighting the necessity for explainable AI. Interpretable models (like decision trees or linear regression) are inherently understandable by humans and allow forensic experts to trace outcomes through transparent logic.[11]

Explainable AI bridges the two worlds by using methods that generate human-understandable explanations of black box output, which is usually done by techniques like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) can highlight which input features contributed most to a prediction, helping forensic analysts justify results. As such, XAI is particularly valuable in contexts where both high accuracy and legal admissibility are required. [12]

3.4 Legal and ethical importance of AI transparency in court evidence

In the forensic context, handling digital evidence requires both technical rigor and legal defensibility. ISO/IEC 27037 provides internationally accepted guidelines for the identification, collection, acquisition, and preservation of digital evidence, emphasizing chain of custody and data integrity. NIST publications, particularly the *Four Principles of Explainable AI* (NIST IR 8312), mandate that AI systems: (1) supply explanations for outputs, (2) present those explanations in a meaningful way to users, (3) maintain accuracy of explanations, and (4) recognize their limitations when operating outside expected conditions [13]. Legally, forensic methods must also meet admissibility criteria: **Frye standard** demands that techniques are "generally accepted" in their field,[14] while **Daubert** requires methods to be testable, peer-reviewed, with known error rates, and transparent.[15] Black-box AI models, which obscure internal logic, may fail these standards, whereas explainable or interpretable AI aligns well with NIST guidelines and supports legal requirements—making it far more suitable for forensic use.

3.5 File Carving Process using "Smart Carving"

The recovery of fragmented files typically follows a **three-phase process**—preprocessing, collation, and reassembly as seen in figure 2—with **Smart Carving** extending it to include a **post-processing step** for enhanced precision [6].

1. **Preprocessing**: Allocated clusters are filtered out using file system metadata, reducing the volume of data for analysis. This also helps distinguish active from deleted files. Tools like *The Sleuth Kit* implement this phase to **extract unallocated space** from disk images.
2. **Collation**:

   Remaining unallocated blocks are grouped by file type using signature analysis, keyword detection, entropy comparison, or byte-pattern matching. More advanced techniques such as **Normalized Compression Distance** (It compares how much extra space is needed to compress two files together versus compressing them separately) and statistical analysis help detect file similarities and boundaries.

   Entropy, in this context, refers to the measure of randomness within a file fragment. High entropy typically indicates compressed or encrypted data, while low entropy suggests structured or human-readable content, like plain text.
3. **Reassembly**: This stage reconstructs files by reordering fragments. Algorithms like **Sequential Hypothesis Testing (SHT)** (data is evaluated as it is collected, rather than waiting for a full dataset before making a decision.) evaluate whether adjacent blocks belong to the same file, while **Parallel Unique Path (PUP)** identifies new fragment starting points. Enhancements like **close region sweep** decode areas near known headers, improving recovery accuracy—especially for complex formats like PDFs or multimedia files. Tools like *ReviveIt* implement such methods effectively.

This structured process allows forensic analysts to recover files more efficiently, even under heavy fragmentation, making it particularly valuable in investigations involving deleted or tampered digital evidence.

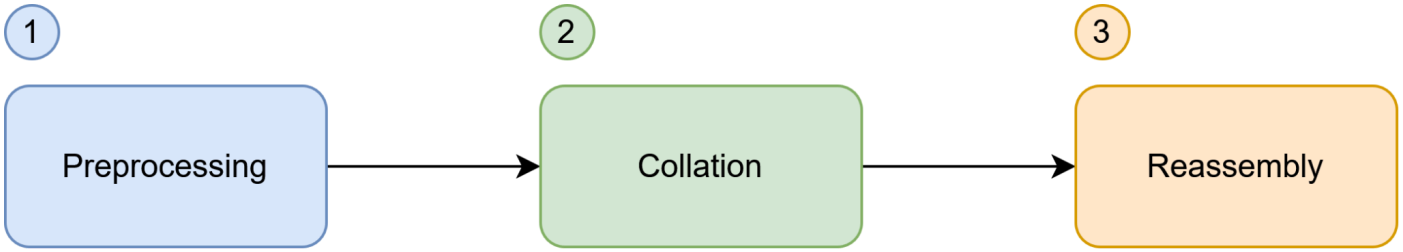


Figure 2: *File Carving Workflow Diagram.*

4. Methodology

This research aims to explore explainable artificial intelligence (XAI) approaches in the classification of image file fragments, a critical task in digital forensics, particularly for recovering and verifying fragmented files in sensitive cases. The methodology focuses on the structure and logic of the SIFT (Sifting File Types) approach through theoretical design and simulated outputs rather than full implementation [16].

4.1 Dataset Structure

The dataset used in this research is based on the publicly available Garfinkel file fragment corpus [18], particularly the GovDocs1 dataset, which has been extensively used in digital forensics studies.

In the context of the SIFT paper, a subset was extracted to include:

- 47,482 file fragments, each exactly 512 bytes in length
- Fragments were sampled from 20 different file types, including but not limited to .jpg, .pdf, .png, .doc, .xls, and .gz
- Fragments are randomly sampled from both the header, body, and footer sections of full files to simulate real-world partial data found in disk or memory forensics
- Each fragment is labeled according to its original file type, enabling **supervised learning**

The original dataset contains over 1 million real files from US government websites, captured and organized by file type (e.g., govdocs1/zip/, govdocs1/pdf/, etc.). Due to the inclusion of active malware in some of the original files, special handling and sandboxed environments are recommended for direct use.

As this research focuses on classification of file types from binary fragments, only the fragment-level metadata (type, size, offset) and content were used for training and testing. No semantic or textual content was processed beyond the raw bytes.

4.2 Feature Extraction

In line with the methodology presented in the SIFT: Sifting file types—application of explainable artificial intelligence in cyber forensics paper, feature extraction was performed using statistical and structural characteristics of the binary file fragments.

Preprocessing:

SIFT preprocesses files by excluding those with size less than twice the fragment size, removes duplicates, and extracts fragments at the byte level, ensuring all fragments are of equal size. The set of raw fragments is extracted from a dataset, and an extra byte is added to each fragment to use as the Class (file type) label

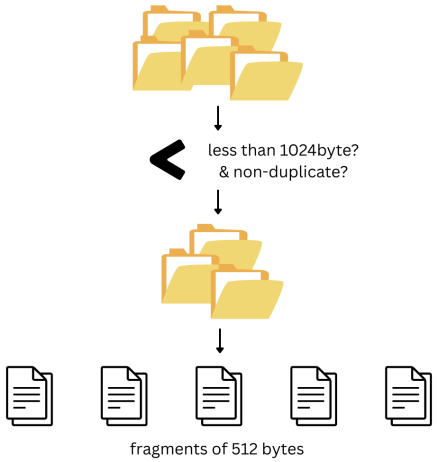


Figure 3: file fragmentation and filtration

Feature Selection:

In accordance with the methodology proposed by the SIFT framework, each file fragment—of fixed size 512 bytes—is treated as a sequence of "terms", where each unique byte value in the range 0 to 255 represents a distinct term. To quantify the relevance of each byte within a fragment relative to its distribution across the entire corpus, Term Frequency–Inverse Document Frequency (TF-IDF) is employed.

- **Term Frequency (TF)** measures how often a specific byte occurs within an individual fragment.
- **Inverse Document Frequency (IDF)** evaluates how distinctive that byte is across all fragments in the dataset, assigning higher values to bytes that appear in fewer fragments and lower values to those that are common.

The TF-IDF score for each byte is computed as:

$$TF_j = \frac{fb_j}{R} \quad and \quad IDF_j = log\left(\frac{\sum_{n=1}^{N} m_n}{K_j}\right)$$

Figure 4: TF and IDF equations

where, fbj denotes the frequency of byte bj within a fragment f, and Kj represents the total number of fragments containing the byte bj. Mn is the number of fragments (documents) containing byte **bj,** and R is is the total number of bytes in the fragment (often the fragment size).

Based on these definitions, we assign weight to a byte b j as follows:

$$W_j = TF_j \times IDF_j$$

Figure 5: Weight equation for each byte

We build a vector of the fragments with selected features FS in the form of a matrix as follows

$$FS = \bigcup_{i=1}^{N} \bigcup_{j=1}^{S+1} \begin{cases} W_j & \text{if } W_j > 0 \\ 0 & \text{otherwise} \\ i & \text{if } j = S+1 \end{cases}$$

Figure 6: build vector matrix with selected features

This results in a 256-dimensional feature vector for each fragment. As part of the feature selection process, byte positions with zero TF-IDF weight are discarded. Consequently, only byte values with meaningful discriminative power are retained for training the classification model, improving both efficiency and explainability.

4.3 Model Design

The core of the SIFT model lies in choosing interpretable classifiers that maintain high accuracy while offering insights into how decisions are made. Initially, Decision Trees were considered because they provide a clear structure showing how features influence classification. However, Random Forest was ultimately preferred due to its superior accuracy and ensemble robustness while still being relatively interpretable [16].

Random Forest is a machine learning method that works by combining the results of many simple decision trees. Each tree makes a guess, and the model chooses the most common answer among them. This makes the predictions more accurate and less likely to be wrong, while still allowing us to understand how decisions were made by looking at the individual trees.

The full classification pipeline includes:

- **Preprocessing**: Filtering small or duplicate files and normalizing byte sequences
- **Vectorization**: Using TF-IDF to convert fragments into byte-weighted feature vectors
- **Feature Relevance Scoring**: Identifying important bytes using LIME and SHAP techniques
- **Classification**: Training and testing models such as Decision Tree, Random Forest, and MLP

A **Multilayer Perceptron** (MLP) is a type of deep learning model composed of multiple layers of neurons that can learn complex patterns in data. However, because of their internal complexity, MLPs are considered "black box" models, meaning their decision-making process is not easily understandable by humans.

While deep models like the Multilayer Perceptron (MLP) were explored—with five hidden layers, and a number of neurons possibly based on the number of file types—they were mainly used for benchmarking. Their 'black-box' nature made them less suitable for forensic applications, where transparency is critical [15].

Random Forest offered the best compromise: it achieved excellent accuracy and allowed for individual decision trees to be interpreted. Each tree in the forest uses a subset of features, and the ensemble votes on the file type, reducing overfitting while maintaining explainability.

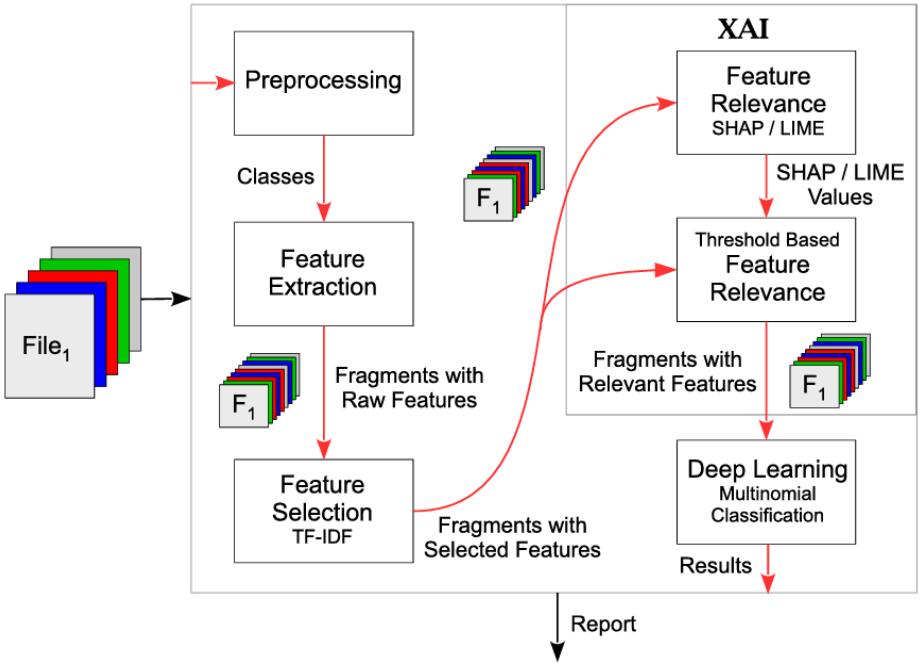
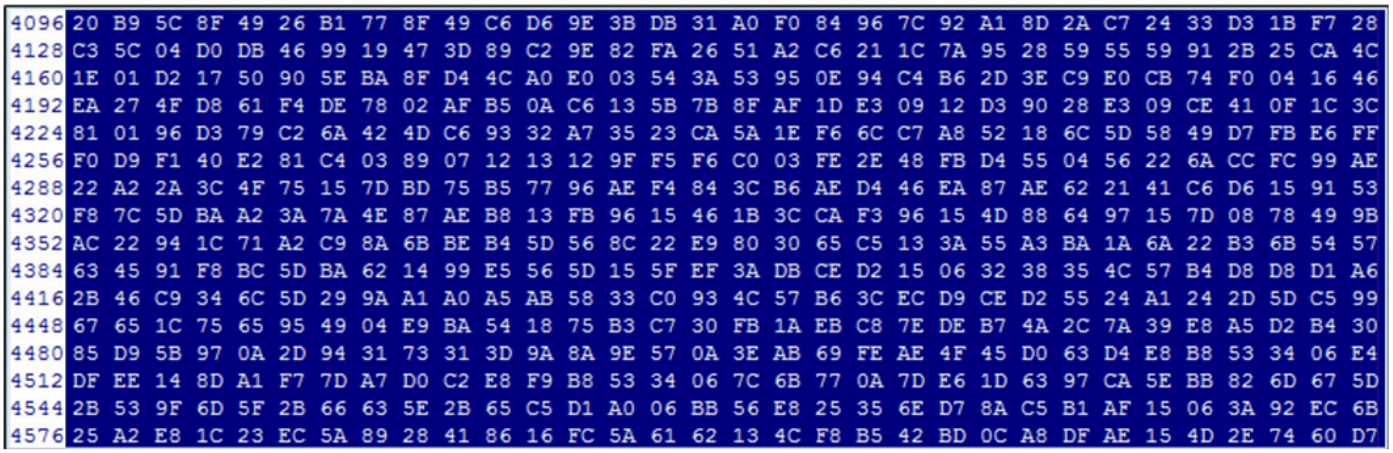


Figure 7: *Overview of SIFT system*



4096	20	B9	5C	8F	49	26	B1	77	8F	49	C6	D6	9E	3B	DB	31	A0	F0	84	96	7C	92	A1	8D	2A	C7	24	33	D3	1B	F7	28
4128	C3	5C	04	D0	DB	46	99	19	47	3D	89	C2	9E	82	FA	26	51	A2	C6	21	1C	7A	95	28	59	55	59	91	2B	25	CA	4C
4160	1E	01	D2	17	50	90	5E	BA	8F	D4	4C	A0	E0	03	54	3A	53	95	0E	94	C4	B6	2D	3E	C9	E0	CB	74	F0	04	16	46
4192	EA	27	4F	D8	61	F4	DE	78	02	AF	B5	0A	C6	13	5B	7B	8F	AF	1D	E3	09	12	D3	90	28	E3	09	CE	41	0F	1C	3C
4224	81	01	96	D3	79	C2	6A	42	4D	C6	93	32	A7	35	23	CA	5A	1E	F6	6C	C7	A8	52	18	6C	5D	58	49	D7	FB	E6	FF
4256	F0	D9	F1	40	E2	81	C4	03	89	07	12	13	12	9F	F5	F6	C0	03	FE	2E	48	FB	D4	55	04	56	22	6A	CC	FC	99	AE
4288	22	A2	2A	3C	4F	75	15	7D	BD	75	B5	77	96	AE	F4	84	3C	B6	AE	D4	46	EA	87	AE	62	21	41	C6	D6	15	91	53
4320	F8	7C	5D	BA	A2	3A	7A	4E	87	AE	B8	13	FB	96	15	46	1B	3C	CA	F3	96	15	4D	88	64	97	15	7D	08	78	49	9B
4352	AC	22	94	1C	71	A2	C9	8A	6B	BE	B4	5D	56	8C	22	E9	80	30	65	C5	13	3A	55	A3	BA	1A	6A	22	B3	6B	54	57
4384	63	45	91	F8	BC	5D	BA	62	14	99	E5	56	5D	15	5F	EF	3A	DB	CE	D2	15	06	32	38	35	4C	57	B4	D8	D8	D1	A6
4416	2B	46	C9	34	6C	5D	29	9A	A1	A0	A5	AB	58	33	C0	93	4C	57	B6	3C	EC	D9	CE	D2	55	24	A1	24	2D	5D	C5	99
4448	67	65	1C	75	65	95	49	04	E9	BA	54	18	75	B3	C7	30	FB	1A	EB	C8	7E	DE	B7	4A	2C	7A	39	E8	A5	D2	B4	30
4480	85	D9	5B	97	0A	2D	94	31	73	31	3D	9A	8A	9E	57	0A	3E	AB	69	FE	AE	4F	45	D0	63	D4	E8	B8	53	34	06	E4
4512	DF	EE	14	8D	A1	F7	7D	A7	D0	C2	E8	F9	B8	53	34	06	7C	6B	77	0A	7D	E6	1D	63	97	CA	5E	BB	82	6D	67	5D
4544	2B	53	9F	6D	5F	2B	66	63	5E	2B	65	C5	D1	A0	06	BB	56	E8	25	35	6E	D7	8A	C5	B1	AF	15	06	3A	92	EC	6B
4576	25	A2	E8	1C	23	EC	5A	89	28	41	86	16	FC	5A	61	62	13	4C	F8	B5	42	BD	0C	A8	DF	AE	15	4D	2E	74	60	D7

Figure 8: *Example of a fragment, of size = 512 bytes, extracted from one of the dataset files used in this paper*

4.4 Explainability

Explainability is a central pillar of this research. Understanding why a model predicted a specific file type is as important as the prediction itself, especially in legal contexts.

Explainable AI (XAI) refers to a set of methods designed to make AI model predictions understandable to humans. This is essential in forensic contexts where legal evidence must be interpretable.

To address this, two leading XAI techniques are applied:

- **LIME (Local Interpretable Model-Agnostic Explanations)**: works by making small changes to an input and observing how the model's prediction changes. This allows it to build a simplified local model to explain what features (bytes) were most important for that specific decision. [16].
- **SHAP (SHapley Additive exPlanations)**: is based on game theory and assigns each input feature a value that reflects its contribution to a prediction. It provides both individual (local) and overall (global) interpretability of model behavior. [16].

These tools make it possible to highlight, for example, that a high frequency of bytes like 0xFF and 0xD8 contributed to classifying a fragment as a JPEG. This kind of byte-level interpretability helps forensic experts validate model decisions and present them as evidence in court.

4.5 Evaluation & Legal Relevance

To assess the effectiveness of the proposed file fragment classification approach, the model was evaluated using standard performance metrics on a labeled dataset of 47,482 file fragments spanning 20 distinct file types.

The dataset was split into training and testing subsets. A variety of classifiers were tested, including: Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Naive Bayes, Logistic Regression.

The primary classifier used in the final model was the Random Forest, chosen for its strong balance between accuracy and interpretability. All models were trained and evaluated using 10-fold cross-validation to ensure generalizability and reduce the likelihood of overfitting.

The evaluation metrics used were:

- Accuracy: The proportion of correctly classified fragments.
- Precision: The proportion of correctly predicted positive observations to the total predicted positives.
- Recall: The proportion of correctly predicted positive observations to all actual positives.
- F1-Score: The harmonic mean of precision and recall, providing a balance between the two.

- Confusion Matrix: To provide detailed insight into the classification performance per file type.

Cross-validation was selected to ensure that the performance metrics reflect the model's ability to generalize to unseen data. Random Forest was preferred for its robustness to noise and its ability to capture complex patterns while remaining explainable to forensic practitioners.

The Random Forest classifier in SIFT achieved:

- Accuracy: 98.82%
- F1-Score: 98.84%
- Precision: 98.92%
- Recall: 98.91%

These results demonstrate that the TF-IDF–based feature representation, combined with explainable machine learning models, can effectively distinguish between file types based on partial binary data. The evaluation confirms the robustness of the proposed method, even when using short fragments of only 512 bytes.

The paper emphasizes the importance of explainability and interpretability of AI-based forensic tools in legal contexts. It discusses how tools like LIME and SHAP can provide transparent explanations of classification results, which are crucial for gaining trust in court proceedings. Specifically, the authors highlight that visualizations and explanations of AI decisions, such as detecting manipulated videos (deepfakes), can serve as trustworthy evidence in a court of law. The use of AI explanations aims to enhance the credibility and transparency of forensic analyses, thereby facilitating their acceptance as admissible evidence in legal processes. Overall, the paper underscores that explainable AI techniques are vital for ensuring that forensic evidence derived from AI models is interpretable, trustworthy, and legally acceptable.

5. Results & Discussion

This research reviewed and analyzed existing approaches to file type classification in the context of digital forensic image carving, with a particular focus on the application of Explainable Artificial Intelligence (XAI).

We examined how traditional file carving methods (header-footer, file structure-based) are being extended with smart carving techniques that incorporate AI models to improve classification accuracy on fragmented or partial data. In this context, we reviewed the SIFT framework as a case study for applying TF-IDF–based feature extraction and interpretable classifiers (e.g., Random Forest) to file fragment classification tasks.

Through this analysis, we observed that:
Interpretable models such as Decision Trees and TF-IDF–based Random Forests can achieve classification accuracy comparable to black-box models (e.g., deep learning), while providing traceable decision paths, which are essential in forensic environments.

Explainability is not only a technical advantage but also a legal necessity in forensic investigations, where evidence must be reproducible and verifiable.

Models like SIFT demonstrate how text-based methods (TF-IDF) can be successfully adapted to binary data to produce both high performance and explainable results in file type classification.

6. Conclusion

This study explored the integration of Explainable AI techniques into the process of file type classification within the broader context of digital forensic image carving. By analyzing smart carving methods and frameworks such as SIFT, we highlighted the benefits of using interpretable models for classifying partial file fragments.

The findings underscore that explainability in AI is not merely a desirable feature, but a **practical and legal requirement** in forensic science, where transparency, accountability, and reproducibility are paramount. While black-box models may offer slightly higher accuracy, they fail to meet the evidentiary standards required in legal settings.

This work contributes to the field by synthesizing technical methodologies and legal standards, advocating for a shift towards **interpretable, domain-aligned AI models** in digital forensics. Future work could involve implementing and comparing such models on real forensic images, integrating entropy-based selection, or enhancing smart carving with hybrid approaches that combine pattern recognition with rule-based logic.

7. Future Work

This research opens several avenues for future exploration:

1. **Practical Implementation**
   Future studies can implement the TF-IDF–based models on real forensic disk images to validate the theoretical findings through empirical results.
2. **Dataset Expansion and Customization**
   Creating a controlled dataset of file fragments with known ground truth would enable more rigorous testing, especially for encrypted or compressed files.
3. **Hybrid Carving Techniques**
   Combining smart carving with structural rules (e.g., signatures or entropy checks) may improve classification when AI predictions are uncertain or insufficient.
4. **Model Comparisons**
   Comparing interpretable models with black-box models in terms of accuracy, explainability, and legal acceptance could guide their application in forensic contexts.
5. **XAI Visualization Tools**
   Developing tools that visualize model decisions would help forensic analysts better understand and defend AI-driven findings in court.
6. **Integration with Forensic Suites**
   Incorporating explainable models into tools like Autopsy or FTK could enhance usability and promote explainability in real investigations.

8. References

[1] S. Parkinson and S. Khan, "The Role of Artificial Intelligence in Digital Forensics: Case Studies and Future Directions," Ai in DF _ Case Studies and Future Directions, 2023.

[2] P. Malik et al., "Enhancing Forensic Analysis of Digital Evidence Using Machine Learning: Techniques, Applications, and Challenges," International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences, vol. 12, no. 5, Sep. 2024.

[3] A. A. Solanke, "Explainable Digital Forensics AI: Towards Mitigating Distrust in AI-Based Digital Forensics Analysis Using Interpretable Models," Forensic Science International: Digital Investigation, vol. 42, 2022.

[4] A. A. Solanke and M. A. Biasiotti, "Digital Forensics AI: Evaluating, Standardizing and Optimizing Digital Evidence Mining Techniques," KI - Künstliche Intelligenz, vol. 36, pp. 143–161, 2022.

[5] D. Dunsin et al., "A Comprehensive Analysis of the Role of Artificial Intelligence and Machine Learning in Modern Digital Forensics and Incident Response," Forensic Science International: Digital Investigation, vol. 48, 2024.

[6] EC-Council, Forensic File Carving: A Guide to Recovering Critical Digital Evidence, EC-Council Whitepaper, 2020.

[7] Roussev, V., & Richard, G. G. (2005). Breaking the Performance Wall: The Case for Distributed Digital Forensics. Digital Investigation.

[8] Stamus Networks Blog – “File Type Spoofing: When a PNG Isn’t Really a PNG”

[9] ADF Solutions – “Using AI and Classification for CSAM Detection”

[10] Timan, T., & van Brakel, R. (2016). “Black-boxed politics: Transparency and algorithmic accountability.” Philosophy & Technology.

[11] Solanke, A. A. (2022). Explainable digital forensics AI: Towards mitigating distrust in AI-based digital forensics analysis using interpretable models. In: Proceedings of the Twenty-Second Annual DFRWS USA

[12] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 51(5), Article No. 93.

[13] Phillips, P., Hahn, C., Fontana, P., Yates, A., Greene, K., Broniatowski, D., & Przybocki, M. (2021). Four principles of explainable artificial intelligence. NIST Interagency/Internal Report (NISTIR) 8312. National Institute of Standards and Technology.

[14] Legal Information Institute. (n.d.). Frye standard. Cornell Law School. Retrieved June 18, 2025

[15] Daubert Standard. (n.d.). In Legal Information Institute. Cornell Law School. Retrieved June 20, 2025,

[16] S. Alam and A. K. Demir, “SIFT: Sifting file types—application of explainable artificial intelligence in cyber

forensics,” Cybersecurity, vol. 7, art. no. 52, Sep. 11, 2024. doi: 10.1186/s42400-024-00241-9.

[17] M. Scanlon et al., “ChatGPT for digital forensic investigation: The good, the bad, and the unknown,” Forensic Science International: Digital Investigation, vol. 46, 2023.

[18] DigitalCorpora.org, "File Corpora," DigitalCorpora, [Online]. Available: https://digitalcorpora.org/corpora/file-corpora/files/ . [Accessed: June 21, 2025].