DOI: **10.37943/25IIRG2444**

Leila Rzayeva
PhD, Research and Innovation Center "CyberTech"
l.rzayeva@astanait.edu.kz, orcid.org/0000-0002-3382-4685
Astana IT University, Kazakhstan
Aigerim Alibek
Master of Information Security, senior lecturer of School of Cybersecurity
A.Zhenisbekkyzy@astanait.edu.kz, orcid.org/0009-0002-0869-9066
Astana IT University, Kazakhstan
Perizat Tazhibayeva
Master's student, BSc, researcher,
242924@astanait.edu.kz, orcid.org/0009-0009-1566-636X
Astana IT University, Kazakhstan
Ali Myrzatay
PhD, researcher, Research and Innovation Center "CyberTech"
A.Myrzatay@astanait.edu.kz, orcid.org/0000-0001-8456-1478
Astana IT University, Kazakhstan
Murat Zhakenov
PhD, scientific consultant, Research and Innovation Center "CyberTech"
muratzhakenov@outlook.com, orcid.org/0009-0005-9672-4365
Digital Heritage of Eurasia LLP, Kazakhstan
Astana IT University, Kazakhstan

A NOVEL ACOUSTIC-ASSISTED CHIP-OFF FRAMEWORK FOR DATA EXTRACTION FROM DAMAGED HARD DISK DRIVES

Abstract: This article discusses best practices for extracting data from damaged mobile phones and hard drives while maintaining the integrity of the storage hardware. It emphasizes that data recovery is essential for digital forensics and cybersecurity due to a common approach to data recovery from mobile devices. In many cases, step-by-step low-level collections instead of quick logical groups reveal hidden artifacts or recently deleted files. Sometimes, this is the only reliable option.

Hard disk errors are usually divided into two categories: logical errors and physical damage. The recovery platform combines proven diagnostics, predictive analysis, and specially designed tools, ranging from installing a magnetic head and replacing an image disk to changing file system settings to make the data readable again. One of the new ideas is acoustic perception of the environment. Just as a device listens to the sound of a running engine, it listens to an acoustic response that can be used to automate the detection of mechanical defects. Tics or stuttering can tell you a lot. The study includes two models: one for detecting problems on the hard drive and the other for data recovery. Model A detects errors related to noise, and Model B tries to recover the data. Thus, this study uses a combined approach to extract data from a damaged hard drive.

With the proliferation of devices for the Internet of Things, acoustic-enabled chip disconnection methods provide forensic protection for repairing and inspecting damaged equipment, such as sensors and damaged industrial components. These results should be of interest to research groups, corporate lawyers, and criminologists in terms of broader coverage and reliability of data recovery operations.

Keywords: signal classification; machine learning; forensic analysis; SMART attributes; logical malfunction; environmental sound recognition; chip-off; data recovery.

Received: 03.02.2026 Accepted: 25.02.2026 Published: 30.03.2026

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

Introduction

Hard drives remain the primary means of data storage for digital forensics and data recovery, but they face logical challenges and a real risk of physical harm. These issues often lead to the loss of large amounts of data by individuals, companies, and groups. In addition, during unsuccessful investigations, deliberate measures are increasingly being used to counteract forensic examination, such as data concealment, removal of traces, and destructive devices.

The traditional method to remove a chip is to remove the memory controller to access the memory. This works well when the standard logic doesn't work. However, these methods require careful manual hearing testing to identify the problem. This slows down the process, depends on the abilities of one person, and can lead to mistakes.

Recent advances in acoustic perception of the environment have shown that support for machine learning parameters, especially vector machines, allows for very precise processing of complex audio formats. Chromaticity detection and spectral contrast monitoring are separate from ambient noise. They also choose a small form of sound. All this is an effective tool for automatic error monitoring on hard drives.

As part of these developments, new technologies for deleting data from hard drives have been replaced by automatic data-based error categorization, which offers consistency and speed of diagnosis. consistency and speed of diagnosis. To accurately detect errors, the command creates a set of instructions for acoustic data about the disk status, such as default, idle, and active. To do this, follow these steps: Make changes to specific hardware, including remote monitoring, configuration, interference removal, and magnetic head replacement, into a single chain of reusable solutions. This includes sic data collection, automatic error monitoring, and process recovery. The accuracy of the readings can be maintained even in the event of serious physical damage.

Literature Review

Recovering data from a hard drive remains a major challenge in digital forensics. Traditional data mining methods often fail when physical damage or limited access interferes with standard logical data collection and requires advanced solutions such as chip technology [1, 2]. When the chip is turned off, the memory chip is removed from the hard disk and read directly without using damaged components [4, 5]. This method is effective but requires high accuracy due to errors that lead to irreversible data loss [7, 8]. Recent research on the abandonment of microchips is aimed at increasing efficiency and reducing physical costs when it comes to storage media [3]. It is a chip removal technology using special storage devices that can be accessed directly without soldering or desoldering [6]. Despite the progress made, it is difficult to determine the best recovery strategy, as the problem persists when the cause of the hard disk error is unclear [9, 10]. This change in sound efficiency indicates problems such as headaches [11]. Recovery is possible by providing information about the device's status at the initial stage [12, 13].

Recovery of hard disk drive (hard disk) data remains an important challenge in digital forensics, cybersecurity, and information protection. Traditional extraction approaches often fail when physical damage or limited access interferes with standard logical acquisition and require advanced solutions such as chip technology [1, 2]. Off the chip, the memory chip is physically removed from the disk and read directly, bypassing defective components [4, 5]. This method is effective but requires excellent accuracy. Any procedural errors can lead to irreversible data loss [7, 8]. Recent research on chip-off recovery has focused on improving efficiency and minimizing physical stress on storage media [3]. Technologies range from solder-based chip removal to the use of special readers that can be accessed directly without complete soldering [6]. Despite these developments, it is difficult to determine the optimal recovery strategy because the problem persists when the cause of the hard drive failure is unclear [9, 10]. At the same time, hard disk error diagnosis is increasingly using acoustic analysis to identify mechanical defects in operating sound. These changes in acoustic signatures can indicate problems such as read/write head failure or spindle motor failure [11]. These non-invasive methods can complement the recovery workflow by providing an early insight into the drive state [12, 13].

Acoustic environment perception improves these diagnoses by applying machine learning algorithms, including support vector machines, to classify audio patterns [14]. Extracts properties such as Mel frequency

DOI: 10.37943/25IIRG2444

spectrum coefficients, color difference vectors, and Mel spectrograms, a method that has proven effective for a wide range of sound classification tasks and can be adapted to hard disk fault profiling. By integrating ESR with the Chipoff workflow, you can pre-classify defects and adapt the recovery procedure to the specific condition of the damaged disk. This can improve success rates and reduce the risk of further damage. However, these integrations are not well studied in the existing literature [15, 16].

In addition, audio descriptors (e.g., MFCCs, spectral contrast) as fed high currents such as SVM or k-nearest neighbors (k-NN) [17]. Their performance was limited by the distinctive ability of the extracted characteristics, which could often be influenced by ambient noise, fluctuations in recording conditions, and the inherent variability of sound [16, 17]. These limitations reduced reliability and adaptability in real-world conditions [18].

The introduction of advanced training has allowed for an in-depth study of characteristics and classifications based on raw signals or spectrograms. Rotating neural networks perfectly recognize spatiotemporal patterns on a spectrograph, while long-term and short-term memory networks effectively recognize temporal dependencies. The transition to data-driven modeling has led to the creation of more accurate and scalable systems, which have been demonstrated in areas such as urban acoustics classification and wildlife monitoring [19, 20]. Additional performance improvements have been made to improve generalization and intensity through improved data expansion techniques, such as transmission (refinement of ready-to-use dataset models for specific areas), as well as pitch changes, time distortion, and synthetic audio input. However, the problem still exists. Ambient noise is often characterized by high variability, overlapping events, background noise, and echoes. Device limitations further limit the functionality of real-time applications and embedded systems, and require continuous improvement of functional representations, neural architectures, and learning strategies [20].

In response to these problems, the study proposes an improved system for recovering data from a hard disk without using chips. At the diagnostic stage, the machine learning model automatically classifies the noise associated with the hard disk error. This non-invasive pre-assessment allows you to obtain useful information before physical intervention, improve error detection, and ensure a more efficient and focused chip removal workflow. This does not replace the usual recovery process, but improves diagnostic accuracy, reduces unnecessary manipulation of damaged media, and optimizes the entire recovery process. At the same time, the existing literature demonstrates a clear research gap: chip removal research mainly focuses on hardware-based recovery procedures, but acoustic diagnostics and acoustic classification based on machine learning are usually considered as separate approaches to error detection, rather than as an integration of chip removal and acoustic pre-diagnosis of the problem of damaged hard drives, which is insufficiently studied. insufficiently studied. Unlike traditional approaches, which are mainly based on manual assessment or isolated diagnostic methods, the proposed system combines hardware-level recovery with automatic acoustic interference classification to provide a more reliable, adaptive, and modern data storage strategy.

Methods and Materials

1. Information extraction methods and acoustic-based defect classification

Data recovery from a faulty hard drive requires diagnostic and technical work. The study used a hybrid search approach that combined the technology of physical chip removal with additional stages of acoustic signal analysis. The procedure begins with an initial assessment of the device using visual inspection and electronic diagnostics to assess the degree of physical damage. If the system detects a disk during connection, the recovery process is performed according to a predefined algorithm. Many modern hard drives are equipped with self-monitoring, analysis, and reporting technologies (S-Monitoring, Analysis, and Reporting Technologies) that provide information about early diagnosis. If access is available at the system level, the disks are connected and analyzed using special software. If access fails, hardware diagnostics are performed. This may include removing the protective cover and replacing or repairing internal components, such as assembling the read/write heads, if necessary. Based on the recovery algorithm, the current process consists of three stages (Figure 1).

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

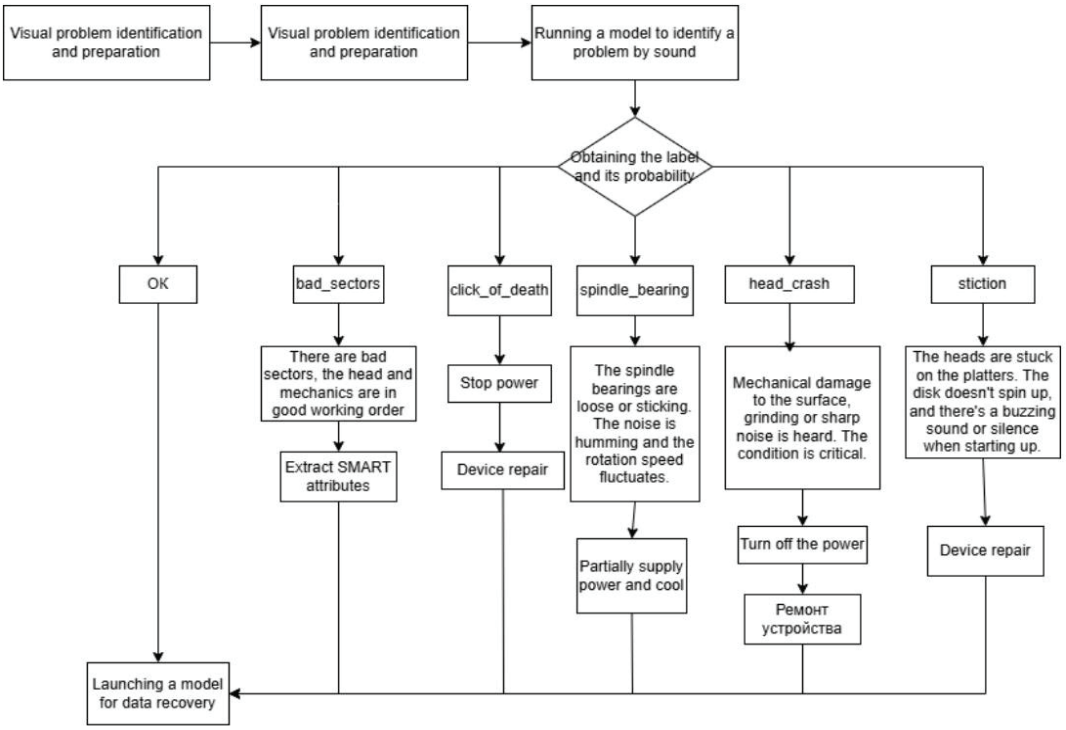


Figure 1. Workflow of the proposed HDD data recovery strategy

An Environmental Sound Recognition (ESR) module has been added to the pre-diagnostic process to improve fault detection at an early stage and limit excessive physical contact with equipment. This system monitors and interprets the acoustic characteristics of a hard drive at all stages of its launch and operation. The key attributes of the signal, including Mel frequency cepstral coefficients (MFCC), Mel spectrograms, and chromaticity vectors, are extracted using the Librosa library and evaluated using the support vector classifier (SVM), which is calibrated to determine specific indicators of mechanical damage.

The results of this acoustic test point to two possible directions:

- If the analysis does not reveal any irregularities, it means that the disk is in good mechanical condition, which allows specialists to continue the recovery process without disassembling the device.
- When certain fault signs are detected, the system can perform chip shutdown procedures with much greater accuracy, generating detailed reports to accurately identify the problem.

This ESR architecture is based on established research methods related to the development of functions and the optimization of classifiers for general acoustic data. In this study, these fundamental concepts are adapted to the unique sound characteristics associated with hard disk failures, which expands the scope of these methods for professional data recovery.

2. Materials

1) Acoustic Fault Classification via ESR

The training on the use of ESR classifiers to diagnose a hard disk was conducted in two stages. A basic vector machine has been developed, which is a Class 10 dataset containing 400 environmental records. This step was very important in testing the reliability of the feature extraction pipeline, which focused on all-in-one, chroma vectors, and mel spectrograms.

To switch to specific diagnostics, a special audio library of hard disk was created. We recorded 36 hard drives (including 11 faulty units) five times each under laboratory settings, capturing everything from power-up to shut-down. To determine when an error occurred, each record was divided into three windows:

- Startup: Covering the initial spin-up and head calibration.
- Idle speed: Monitoring of the drive during passive rotation.
- Boot: Checking the read/write mechanism in simulated voltage mode.

DOI: 10.37943/25IIRG2444

To clarify the structure of the acoustic data set of the hard disk used in this study, some of the material is presented in Table 1. The table varies depending on the initial recording process and the recording used for processing.

Table 1. Structure of the Acoustic HDD Dataset

Dataset stage	Description	Size
Raw acoustic dataset	HDD recordings collected from all drives	180 recordings
Faulty drives subset	Faulty HDDs included in the dataset	11 drives
Processed dataset	Balanced/selected samples used for classification	108 samples
Training set	Samples used for model training	86
Test set	Samples used for testing	22

As shown in Table 1, the experimental dataset was collected in two stages: the collection of raw acoustic recordings from HDDs and the processing of datasets prepared for classification. This structure describes the difference between the full set of sounds recorded from a hard drive and the balanced subset used for training and testing machine learning models.

The final model was created by adjusting the original data on the hard disk based on this new data on the hard disk. Avoid simply memorizing (re-examining) samples and use data extensions to make the model more reliable:

- Stretching the time to change the playback speed.
- Change the pitch to change the frequency.
- Background noise to simulate a messy and real environment.

Finally, using the balanced sampling method, the expanded dataset was divided into 80% for training (86 samples) and 20% for testing (22 samples) to ensure accuracy in all categories.

To process the audio data, we used the Librosa library to extract a variety of specific sound characteristics:

- MFCCs to capture the short-term spectral shape.
- Chroma features to map out pitch-related profiles.
- Mel spectrograms to see how frequency power is distributed.
- Spectral contrast to measure the difference between the "peaks and valleys" of the sound spectrum.
- Tonnetz to provide a six-dimensional map of tonal shifts.
- FFT spectrum to analyze how amplitude spreads across different frequencies.

Extraction was performed in a clear step-by-step mode. This sound was first imported from Librosa. Download and calculate the short-term Fourier transforms. After getting each of the above functions, I combined them into a single data vector. This combined vector was used to form an SVM with a linear kernel for the final classification.

Extraction followed a clear step-by-step workflow. First, audio was imported using librosa

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

$$x = [x_1, x_1 x_2 \dots, x_d] \in R^d \,, \qquad (1)$$

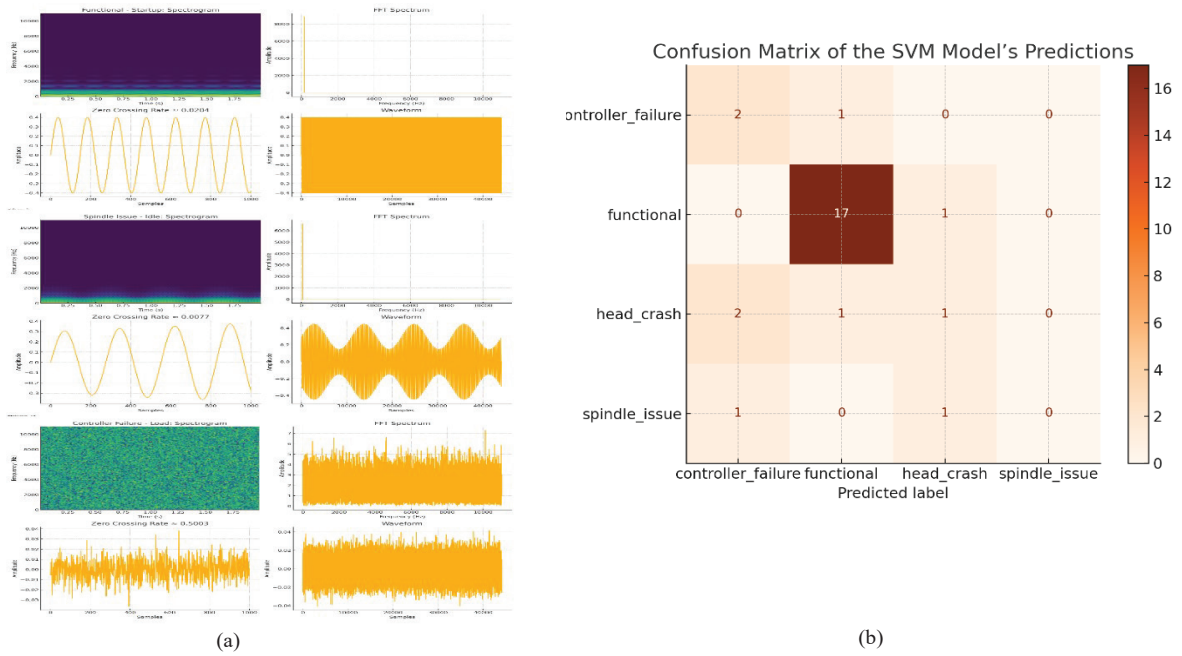
where x denotes the multidimensional descriptor extracted from the HDD sound signal, and d is the total number of acoustic features. In this study, the vector included Mel-frequency cepstral coefficients (MFCCs), chroma features, mel-spectrogram statistics, spectral contrast, tonal centroid features (Tonnetz), and FFT-based descriptors.

Among these, MFCCs were used as one of the main compact descriptors of the spectral envelope. They were computed as

$$MFCC_n = \sum_{m=1}^{M} \quad log(S_m) cos[\frac{\pi n}{M}(m - 0.5)] \,, \qquad (2)$$

where S_m is the energy of the m-th mel-frequency band, M is the number of mel filters, and n is the cepstral coefficient index. This transformation makes it possible to capture frequency-dependent changes caused by spindle instability, head clicks, scraping noise, and other mechanically induced HDD faults.

For a visual comparison, Figure 2 shows the time-frequency and amplitude characteristics of the three scenarios. At the first start, the drives were in good condition, the drives with worn spindles were idling, and the controller was out of order due to high load.



(a) (b)

Figure 2. Acoustic visualizations and classification results used for HDD fault detection: (a) time- and frequency-domain representations of HDD sounds recorded under three operating conditions, including functional startup, spindle-related fault during idle mode, and controller-related fault under load; (b) confusion matrix of the SVM classifier showing the prediction performance for acoustically augmented HDD recordings.

In Fig. 2a, the waveform indicates a change in amplitude over time and helps identify irregular transients associated with abnormal mechanical effects. Spectrograms show how the signal energy is distributed over frequency and time, which makes it possible to identify characteristic acoustic patterns associated with various hard disk failure states. The spectral energy distribution provides additional insight into the prevailing energy range in the frequency domain, which can vary depending on normal and faulty operating modes. Zero transition planning reflects rapid fluctuations in the signal and can help detect changes in the signal structure caused by mechanical instability. Even so, it shows how these acoustic images help

DOI: 10.37943/25IIRG2444

determine the operating conditions of a hard drive. Together, these visualizations improve the interpretability of the diagnostic stage and allow acoustic analysis to pre-classify disk errors prior to physical intervention.

Spectral analysis shows that different operating conditions of the disk lead to very specific energy properties. For example, functional variators have stable frequency characteristics during regular cycles, and spindle problems lead to different energy ranges at high-frequency loads. On the other hand, interference from the controller tends to enhance spectral characteristics with very low background noise. By checking the frequency of transition of sound waves from the very beginning, you can detect anomalies at certain stages of operation. The synthesis of these acoustic markers shows that using ESR analysis for the preliminary classification of disk defects is a very effective approach. To create a classifier, we use the scikit-learn train_test_split function to split the data from our function into 80% for the training group and 20% for the test group. Then, we configure the support vector method using a linear kernel to process the data.

Key steps included:

- Dividing the feature set into an 80:20 split for training and validation.
- Configuring the SVM classifier with a linear kernel for clear categorization.
- Using the fit() method to train the model on the primary dataset.

To classify HDD fault-related acoustic patterns, a Support Vector Machine (SVM) with a linear kernel was used. The decision function of the classifier is defined as

$$f(x) = sign(w^T x + b) \ , \qquad (3)$$

where x idenotes the input feature vector extracted from the HDD acoustic signal, w is the weight vector that determines the orientation of the separating hyperplane, and b is the bias term that shifts the hyperplane in the feature space, and $sign(\cdot)$ is the decision function that assigns the input to one of the target classes. The value $w^T x + b$ determines on which side of the separating hyperplane the (3) sample is located, and the sign of this value defines the predicted class. In this way, the SVM model distinguishes between HDD fault-related acoustic patterns based on their extracted features.

For soft-margin classification, the SVM training objective can be written as

$$min_{w,b,\xi_i} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i$$

subject to

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \dots, N \ , \qquad (4)$$

where N is the number of training samples, $y_i \in \{-1, +1\}$ is the class label, ξ_i are slack variables allowing limited misclassification, and C is the regularization parameter controlling the trade-off between margin maximization and classification error. This formulation is suitable for HDD acoustic diagnostics, where some overlap between normal and faulty sound patterns may occur because of noise, recording conditions, or mixed mechanical symptoms.

3. Evaluating

The initial stage of the physical assessment included a thorough visual inspection of the disk substrate (control board). The purpose was to identify obvious anomalies such as deformed surfaces, surface-mounted parts, or traces of burning, and to verify that the interface connector was damaged. If serious structural defects or burnt-out circuits are detected, the engine is deliberately switched off to avoid electrical or mechanical damage.

The controller board can then be gradually removed to check the contact pad connecting the PCB to the magnetic head assembly (HSA). As shown in Figure 3, these pads are necessary to maintain electrical integrity between the subsystems.

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

Figure 3. Assessment of contact pad integrity after controller board removal. This step verifies electrical continuity before further recovery

If a disk failure starts on the PCB, a simple replacement may not be enough, as each board contains unique "application" data. Access to data often requires a delicate "surgical" transfer of necessary components - like EEPROM, MCU, or NVRAM - from damaged boards to working donor ones. EPPROM is the most important piece of this puzzle. Keep the unique settings of the "application" and the settings necessary to properly configure the disk and read the passport plate. Without this specific data, even the same donor card cannot initialize the disk.

Finding a suitable donor fee requires precision. In this case, it is necessary to make sure that the replacement board exactly meets the specifications of the original, MCU model and VCM/SM controller. Using boards of this shape is very dangerous. Incompatible connections can lead to voltage surges that can damage the charger. When this detonator fails, the disk actually "blinds", turning the replacement of a standard board into an almost impossible recovery operation.

Before assembling and turning on the disc, perform a diagnostic scan using a multimeter. Make sure that the supply line from 5V to 12 volts is closed and measure the resistance of the motor winding. Incorrect readings indicate complete errors that need to be corrected before proceeding. After completing all the motherboard integrity tests, reconnect the drive and start the recovery process.

After powering up the hard drive, as shown in Figure 4, the ESR module is used to check for sound-related problems and detect abnormal noises coming from the disc. If your hard drive is defective, open it and inspect it. If it is really broken, the main magnetic block may need to be replaced. Do this in a clean room using special tools, as even a small particle of dust can cause permanent damage to the plate. Defective heads should be carefully removed, and new ones should be aligned and installed in place without damaging the internal components. During this repair, all operations are performed according to the structured algorithm shown in Figure 1. This ensures accuracy, security, and maximum potential for data recovery.

DOI: 10.37943/25IIRG2444

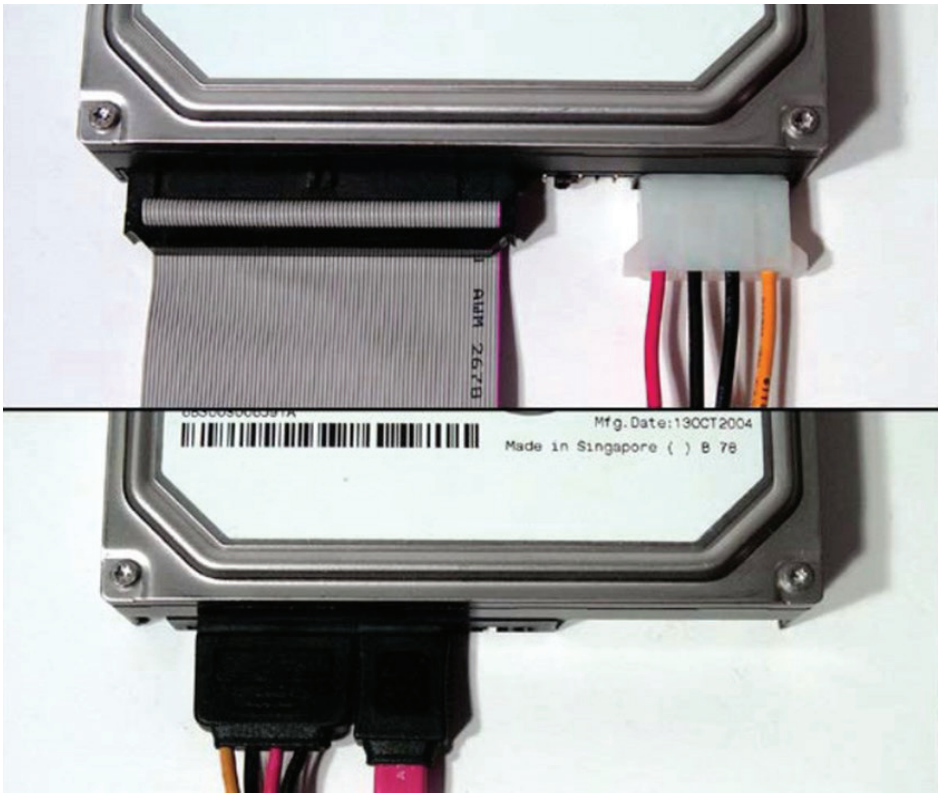


Figure 4. Identification markings on the HDD printed circuit board used to verify device parameters before recovery

Removing the lid opens access to the internal structure of the sealed unit, as shown in Figure 5. At this stage, the upper magnet is carefully separated using a special tool, which allows for further disassembly of the drive system. To ensure stability, a strainer is attached to the magnetic head, after which the damaged block of the magnetic head is removed from the surface of the plate. A similar procedure is performed to obtain a compatible and functional magnetic head unit on a donor disk. After confirmation, the donor head assembly is carefully installed on the target disk for data recovery.

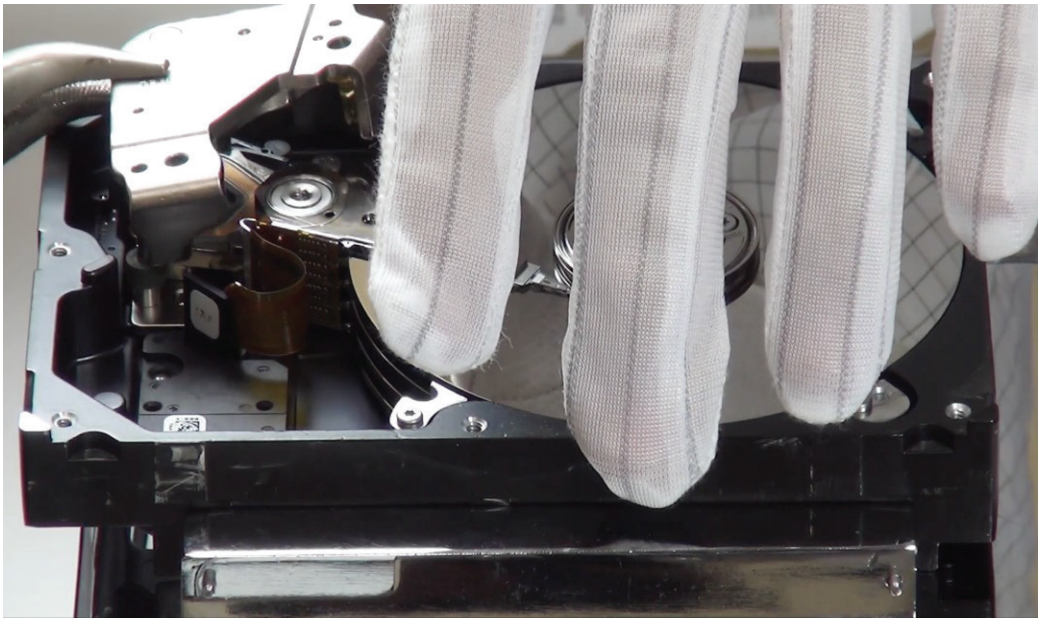


Figure 5. Removal of the upper magnet

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

After the machine is installed, the magnetic head unit is installed in a designated parking space next to the spindle. The upper magnet is installed in place to ensure that the drive is assembled and properly aligned. The sealed case closes with a lid, which completes the replacement of the magnetic head unit. At this stage, the disk is mechanically reassembled and ready for the next stage of diagnostics or data extraction.

After the top magnet was removed, a special hard drive puller was installed as shown in Figure 6.

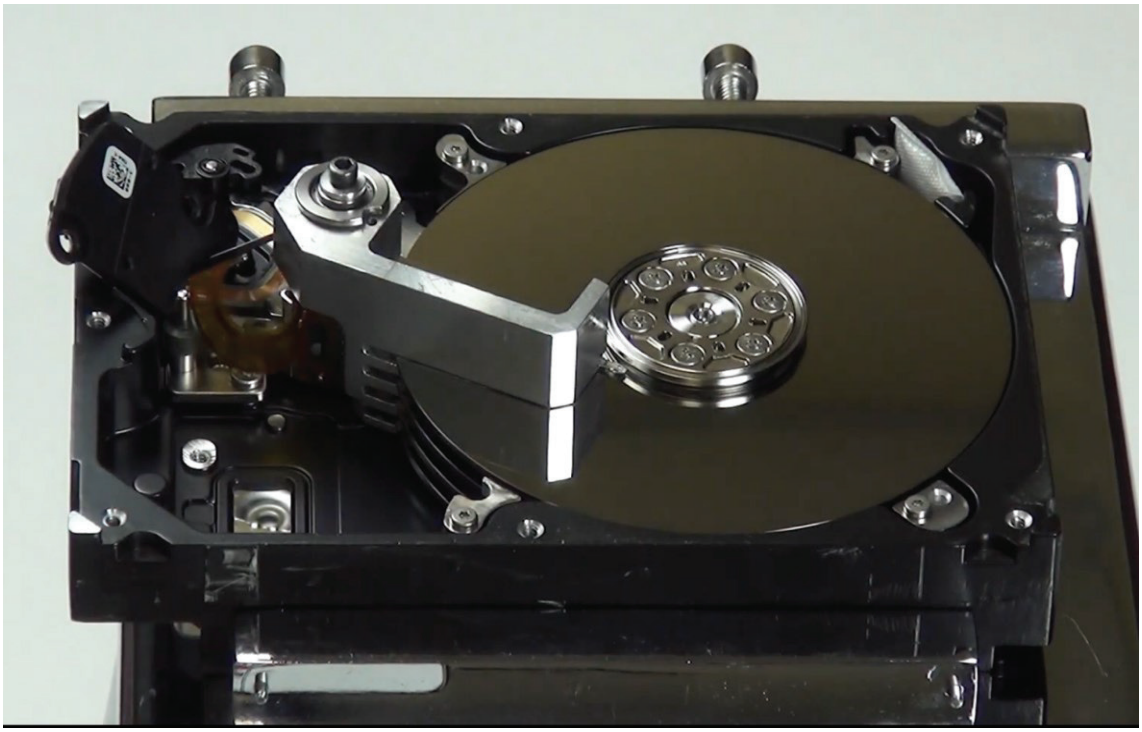


Figure 6. HDD puller used to safely position internal components for the next recovery stage

After the equipment diagnostic procedures and the preliminary error assessment were completed, an assessment was carried out using an experimental verification process, which is used to evaluate the effectiveness of additional acoustic classification steps and general recovery workflows.

Results

Data recovery from a failed hard disk (HDD) is a complex and multifaceted process that requires the use of various methods and techniques. We used a progressive combined approach to data search. First, we performed a thorough diagnostic to determine the extent of the disk damage. Now we can move on to the data extraction stage. An information recovery model is used to extract data from donors. The system is a modular ML platform for file recovery based on content and metadata analysis. Unlike traditional signature-based approaches that use fixed byte patterns, the proposed architecture is prepared using statistical functions and file system metadata. According to the test results, the file processing classifier showed an accuracy of 94.7%, which is 24.7% higher than using the standard magic byte method. Two sets of data were used to prepare and test the developed machine learning-based file recovery system (ML-based File21 Recovery System): real (Laptop Dataset) and synthetic (Synthetic Dataset). The combination of these datasets provides a balance between reliability and diversity of input data, which is important to increase the generalizing ability of the model and its resistance to noise, fragmentation, and corruption of structural files.

Figure 7 shows the ratio of the two datasets depending on the number of files and the total size (in gigabytes).

DOI: 10.37943/25IIRG2444

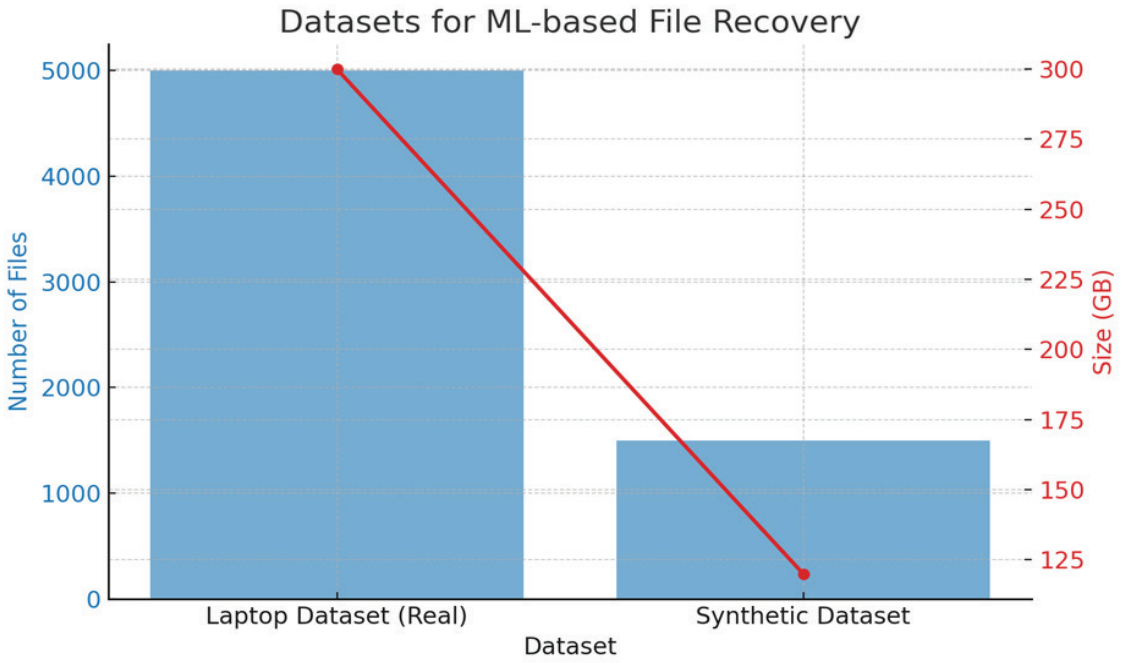


Figure 7. Datasets

To improve the clarity of the dataset description, Table 2 summarizes the composition of the datasets used for file recovery experiments.

Table 2. Summary of dataset composition used for file recovery experiments

Dataset	Number of files	**Total size (GB)**	Purpose
Laptop Dataset	5000	300	Real data for training and validation
Synthetic Dataset	1800	120	Simulation of rare corruption scenarios and dataset expansion
Total	6800	420	Combined experimental dataset

As shown in Table 2, the experimental dataset combines real and synthetic data sources. This structure provides realism and variety, allowing you to evaluate the model in practical cases of file recovery, as well as take into account rare and controlled cases of damage.

Figure 8 shows a high proportion of correct classifications of "clean" formats (JPEG, MP4, PDF, PNG, TXT, ZIP) and a low proportion of corrupted versions (*_CORRUPTED). This is due to the fact that the format features are more pronounced than the corruption features. The most common errors are JPEG_CORRUPTED > JPEG (87 cases), PDF_CORRUPTED > PDF (8 cases), PNG_CORRUPTED > PNG (9 cases). Therefore, the model reliably determines the format but does not distinguish the degree of damage well, as shown in Figure 9.

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

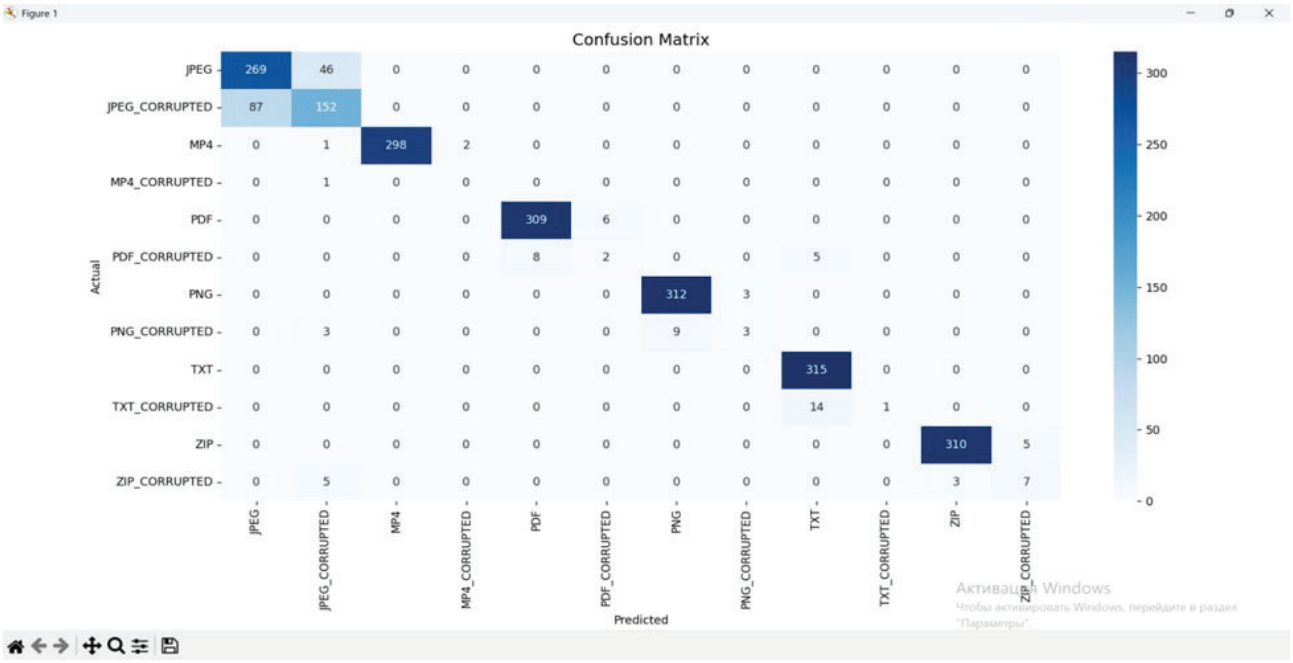


Figure 8. Error matrix of the file recovery model

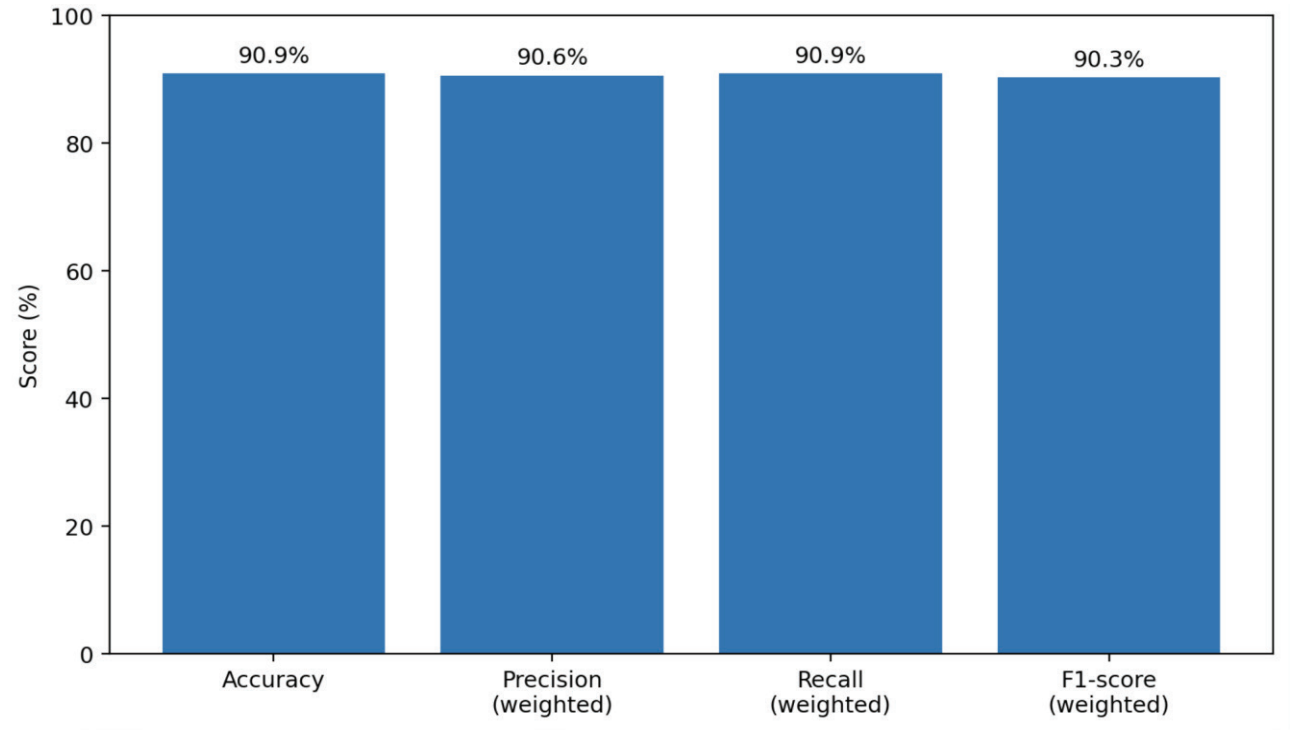


Figure 9. Error matrix of the file recovery model

Figure 10 shows the Receiver Operating Characteristics (ROC) curve of Model B performing data recovery.

DOI: 10.37943/25IIRG2444

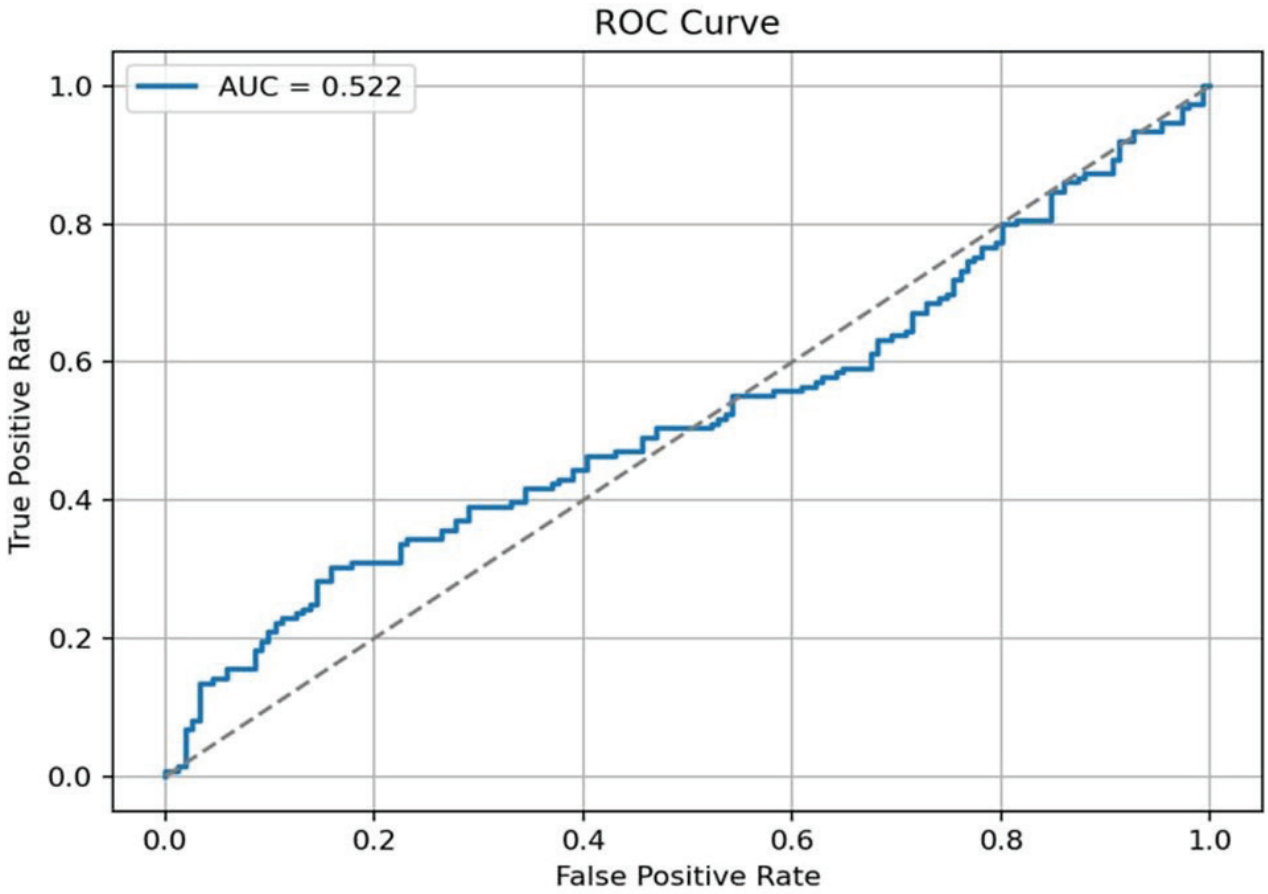


Figure 10. ROC curve of model

The power curve of the Receiver Operating Characteristics (ROC) of Model B, which performs data extraction operations after classifying Model A errors, is shown in Figure 10. The model performs slightly better than the random classification (AUC ≈ 0.5), which corresponds to the area under the curve (AUC = 0.522). This suggests that the current Model B's ability to distinguish between successful and unsuccessful data recovery scenarios is limited. A slight difference is visible on the curves approaching the diagonal. The percentage of false positives increases simultaneously with the percentage of true positives. There are many reasons that can explain this trend. Firstly, the number of logs currently available is insufficient to explain the differences in successful and unsuccessful recovery scenarios, especially if there are fragmentation, interference, or data on a damaged disk. Secondly, since recovery-related signals are affected by incomplete downloads, damaged sectors, computational instability, and structural artifacts that limit class separation, an expanded set of functions can cause significant interference. Thirdly, additional settings are needed to improve generalization and reduce sensitivity to input noise, since the hyperparametric configuration of the existing model is not yet ideal for solving this problem. Given the record-breaking expansion, functional design, and optimization of the model, the significance of the auction result should be seen as proof that the relationship model is at an intermediate stage of development and requires further refinement. It can be used in combination with Model A to probabilistically evaluate the success of data retrieval, and in more complex situations, it can be adaptively redirected to manual or hybrid (human-machine) search methods.

Figure 11 shows the Precision–Recall curve for Model B, which performs the data recovery task after diagnosing faults classified by Model A.

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

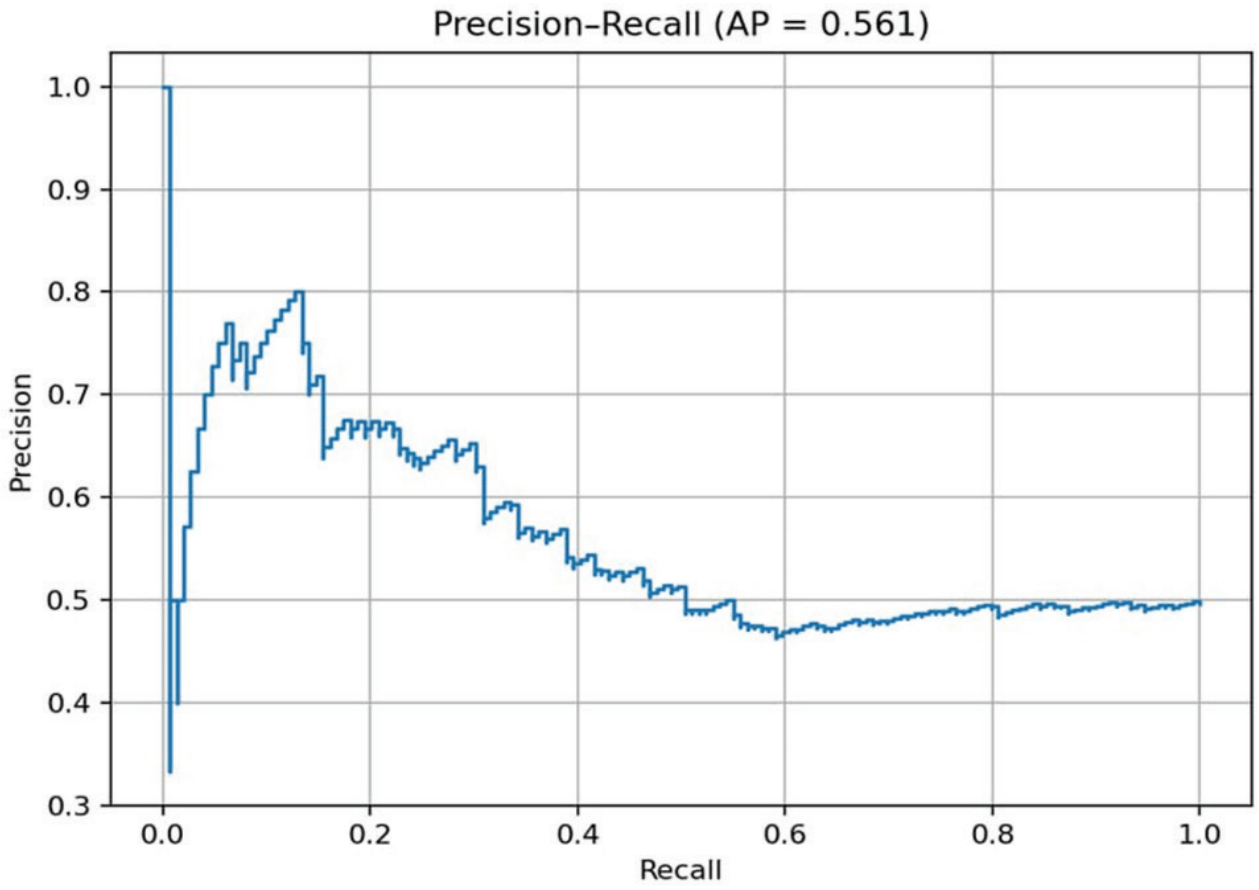


Figure 11. Precision–Recall curve for model B

The abscissa axis shows Recall, which reflects the proportion of successfully recovered cases among all potentially recoverable ones, and the ordinate axis shows Precision, which characterizes the proportion of correct recoveries among all attempts. The average area under the curve (AP = 0.561) indicates a moderate level of agreement between Precision and Recall, which corresponds to a moderate ability of Model B to distinguish between successful and unsuccessful data recovery scenarios. The initial section of the curve (at low Recall values) demonstrates a high Precision level of ≈ 0.8–1.0, which means that the model confidently identifies a portion of cases with a high probability of successful recovery. However, as Recall increases, Precision gradually decreases to the range of 0.45–0.5, reflecting an increase in the number of false positives as the case coverage expands. This dynamic is typical for systems where the data recovery signal is highly noisy, and the model operates under conditions of high uncertainty, such as incomplete dumps or a significant number of damaged sectors. From a practical perspective, this means that the model effectively identifies "reliable" recovery cases but struggles with borderline situations. Improving quality requires further data balancing, increasing the number of positive examples, and implementing multimodal features that consider not only binary recovery results but also error correction metrics, the entropy of the recovered block, and read time. Thus, the graph demonstrates that Model B is at an intermediate stage of maturity: it has a basic ability to filter out successful recoveries but requires further training and integration with the outputs of Model A to achieve a sustainable level of accuracy required for autonomous operation in an intelligent HDD recovery system.

Discussion

Unlike conventional forensic data recovery systems, the system mentioned above combines intelligent acoustic diagnostics with hardware-level processing. Traditional forensic techniques are often based on expert error assessment, replacement of donor parts, and manual physical examination. This procedure often

DOI: 10.37943/25IIRG2444

requires individual experience and frequent use of portable devices. Despite the effectiveness of these traditional operations, there is no complex diagnostic automation system. Thus, the proposed methodology makes it possible to classify defects in advance and better plan strategic repair of microcircuits using acoustic analysis based on machine learning before physical intervention. Ultimately, this simplifies the preparation for recovery and minimizes unnecessary processing of weak media.

However, the current approach has a number of significant drawbacks. Firstly, not all failure scenarios that may arise during a real judicial investigation fit into a limited set of acoustic data. Secondly, the stability and reproducibility of error classification can be affected by the strong dependence of acoustic characteristics on device characteristics, background noise, and recording settings. Thirdly, the Rock Oak results show that the recovery prediction component works at a level comparable to randomized classification, despite the excellent accuracy of determining the type in the file recovery classifier. This means that the model configuration and the current representation of the object are not yet reliable enough to predict offline recovery. Furthermore, since functions such as financial intervention, donor selection, and interpretation of international events always require human involvement, the system aims to support the decision-making process without involving specialists in the field of forensic medicine.

Conclusion

Forensic scenarios involving hard drives with simultaneous physical and logical defects are the subject of a unique chip-off technology in this study to locate HDD data.

The main contributions of this study are:

- It includes an ESR-based acoustic diagnosis for non-invasive detection of the first errors in the hard disk recovery process.
- It develops structured recovery processes for damaged hard drives and microchips that facilitate forensic analysis of devices.
- It develops and evaluates machine learning-based file recovery models with higher classification accuracy than traditional signature methods.

The basic approach simplifies the usual chip removal procedures by combining customer-specific recovery processes with precise machine control. When traditional software tools are no longer useful, you can work directly with the memory components. This method provides reliable recovery of severely damaged equipment and ensures data integrity during the recovery process through the use of professional substrate transfer devices (controller boards) and switching head assemblies. The inclusion of peripheral modules for speech recognition (ESR) is a characteristic novelty in this system. For the analysis of audible indicators with machine learning models, in particular Support Vector Machines (SVM), this component provides a non-invasive diagnostic layer. Based on the acoustic performance, the module serves as a guide to automatically identify mechanical defects and optimize subsequent actions to turn off the chip. The adaptability of this study makes it so important. By adapting the removal technology to the unique condition of the hard drive, the probability of success increases, and the possibility of additional damage during disassembly decreases. However, there are some disadvantages to the proposed chip-off strategy. This must be done in a controlled environment, such as a clean room, and requires expertise. In addition, the width of the training data and the quality of the audio input have a significant impact on the reliability of this module. For example, background noise in the actual configuration can affect the accuracy of the diagnosis. Future studies should continue to include these elements in order to strengthen the system and make it more accessible to the public. This methodology is of fundamental importance for the support of forensic examinations and data recovery of physically damaged or damaged border devices in the context of Internet security. Ensuring data integrity and recovery is becoming a key necessity to ensure digital trust and effective forensic training, as distributed storage architectures serve smart homes, industrial applications, and critical infrastructures. The automation of chipping processes is the subject of future research in order to reduce the need for highly qualified workers. In addition, diagnostic accuracy is significantly improved by adding a wider range of hard disk models and other defect mechanisms to the acoustic data set. Of course, it is also possible to apply this scheme to other storage formats, such as SSD. Finally, our study complements the latest developments in the fields of data

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

recovery and digital forensics. We are building a scalable and automated system to solve the complex data loss problems in the digital age by applying an efficient chip disconnect solution based on national data.

Acknowledgment

This study was conducted with financial support from 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, as part of the scientific project IRN BR24993232 titled "Development of innovative technologies for conducting digital forensic investigations using intelligent software-hardware complexes."

References

[1] S. Tanenbaum and H. Bos, *Modern Operating Systems*, 5th ed. Harlow, U.K.: Pearson Education, 2022.

[2] L. Rzayeva, A. Imanberdi, I. Opirskyy, O. Harasymchuk, and G. Abitova, "Analysis of technical features of data encryption implementation on SD cards in the Android system," *Scientific Journal of Astana IT University*, pp. 157–171, 2025. https://doi.org/10.37943/21LMQF2486

[3] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015. https://doi.org/10.1109/MSP.2014.2326181

[4] P. Cruickshank, "Discarded laptop yields revelations on network behind Brussels, Paris attacks," *CNN*, Jan. 25, 2017. [Online]. Available: https://edition.cnn.com/2017/01/24/europe/brussels-laptop-revelations/index.html. [Accessed: Mar. 27, 2025].

[5] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. 2015 IEEE 25th Int. Workshop Machine Learning for Signal Processing (MLSP)*, 2015, pp. 1–6. https://doi.org/10.1109/MLSP.2015.7324337

[6] "San Bernardino shooters tried to destroy phones, hard drives, sources say," *ABC News*. [Online]. Available: https://abcnews.go.com/US/san-bernardino-shooters-destroy-phones-hard-drives-sources/story?id=35570286. [Accessed: Mar. 27, 2025].

[7] Y. Saraçlıoğlu, B. Saoud, I. Shayea, G. Y. Pil, and L. Rzayeva, "Environmental Sound Recognition (ESR) with Python," in Proceedings - 29th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2025-Summer), 2025, https://doi.org/10.1109/SNPD65828.2025.11254371

[8] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. 2017 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135. https://doi.org/10.48550/arXiv.1609.09430

[9] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017. https://doi.org/10.48550/arXiv.1608.04363

[10] M. Gül and E. Kugu, "A survey on anti-forensics techniques," *International Artificial Intelligence and Data Processing Symposium (IDAP)*, 2017. https://doi.org/10.1109/IDAP.2017.8090341

[11] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1015–1018. https://doi.org/10.1145/2733373.2806390

[12] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2880–2894, 2020. https://doi.org/10.48550/arXiv.1912.10211

[13] S. Schneider, A. Baevski, R. Collobert, M. Auli, and A. Mohamed, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Interspeech*, 2021, pp. 3652–3656. https://doi.org/10.48550/arXiv.2006.11477

[14] J.-P. Van Belle, "Anti-forensics: A practitioner perspective," *International Journal of Cyber-Security and Digital Forensics*, vol. 4, no. 2, pp. 390–403, 2015, https://doi.org/10.17781/P001593

[15] J. Oh and H. Hwang, "Advanced forensic recovery of deleted file data in F2FS," Forensic Science International: Digital Investigation, vol. 54, Art. no. 301976, 2025, https://doi.org/10.1016/j.fsidi.2025.301976

DOI: 10.37943/25IIRG2444

[16] R. Xu, X. Wang, and J. Wu, “Classification based hard disk drive failure prediction: Methodologies, performance evaluation and comparison,” in *Proc. 2022 IEEE 18th Int. Conf. Automation Science and Engineering (CASE)*, Aug. 2022, pp. 189–195, https://doi.org/10.1109/CASE49997.2022.9926720

[17] A. Silberschatz, P. B. Galvin, and G. Gagne, Operating System Concepts, 10th ed. Hoboken, NJ, USA: Wiley, 2018.

[18] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” in *Proc. 2016 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 351–355. https://doi.org/10.3390/app6060162

[19] R. Chandramouli and E. Hibbard, Guidelines for Media Sanitization, NIST SP 800-88 Rev. 2. Gaithersburg, MD, USA: National Institute of Standards and Technology, Sep. 2025, https://doi.org/10.6028/NIST.SP.800-88r2

[20] Scientific Working Group on Digital Evidence. Best Practices for Data Destruction Media Sterilization and Sanitization. – SWGDE F-24-001-1.0, version 1.1. – 2025. – URL: https://swgde.org/wp-content/uploads/2025/11/Best-Practices-for-Data-Destruction-Media-Sterilization-and-Sanitization-24-F-001-1.0.pdf